2018

A methodology for sorting haploid and diploid corn seed using terahertz time domain spectroscopy and machine learning

Jared Lee Taylor
Iowa State University

A methodology for sorting haploid and diploid corn seed using terahertz time domain spectroscopy and machine learning

by

Jared Lee Taylor

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Aerospace Engineering

Program of Study Committee:
C. Thomas Chiou, Co-major Professor
Leonard Bond, Co-major Professor
David A. Laird

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this thesis. The Graduate College will ensure this thesis is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2018

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

Special thanks go to my family. My wife Sarah, daughter Elena, and son Micah were a source of tremendous encouragement and motivation when research was hard.

Thanks belong to Dr. Chiou for his guidance, patience, and support throughout the research and writing process. I would also like to thank my committee members for their time and effort for this work: Dr. Leonard Bond and Dr. David Laird.

Thanks also to Austin Leeds, without his IT knowledge and equipment we may still be training the machine learning models.

ABSTRACT

Terahertz technology has been rapidly expanding both in its use and in attention given to it. A possible application is in corn breeding, specifically when the doubled haploid method is used. Haploid kernels are induced in corn plants in order to decrease the time to reach homozygous genetic corn lines. These haploid kernels must be separated from the surrounding diploid kernels; presently this is done by extensive manual labor using visual markers. This work represents a proof of concept that haploid classification can be automated using terahertz time domain spectroscopy (THz-TDS) paired with a machine learning algorithm, like a probabilistic neural network (PNN).

In this work, a THz-TDS system was used to collect time domain waveforms from a sample of mixed haploid and diploid corn kernels. Variabilities in beam focus and kernel geometry were reduced by taking multiple scans at different heights and at many scan positions. A watershed image segmentation technique was used to reduce the data quantity and organize them by kernel. The waveform data were then transformed to the frequency domain and further classified by PNN with a training set random subsampling technique. Leave-one-out and K-folds cross-validation procedures were used to train the model. The preliminary results show promise yielding an average classification rate of 75 percent correct by 5-fold cross-validation. THz ability to penetrate material leads to immense potential for similar applications in nondestructive evaluation, biomed, and agriculture.

CHAPTER 1. INTRODUCTION

1.1 Corn Breeding and Kernel Sorting

This section will introduce the important information relevant to this work, including corn kernel sorting, THz technology, and machine learning.

1.1.1 Corn Breeding

Corn is ubiquitous in this modern world. Food and Agricultural Organization (FAO) of the United Nations (2015) estimated 1.007 billion tonnes of maize was produced in 2015. Corn breeding is a large part of commercial agricultural research. One of the first steps of breeding corn is the production of homozygous candidate lines. In a homozygous kernel, both sets of chromosomes are identical. If the chromosomes are identical, the next inbred generation will be determined and all progeny kernels will be identical. Traditional corn breeding involves many generations of inbreeding. A 99% homozygous corn line can take six to eight generations. Using doubled haploid breeding techniques (DH), the process can be cut down to two to three generations, according to Prasanna (2012). The DH method involves induction of haploid kernels in a genetic stock. Haploid kernels, as opposed to diploid kernels, have only half of the genetic material (one set of chromosomes). These haploids can later be treated with a chemical called colchicine to induce duplication of the chromosomes, producing a homozygous diploid plant in just two generations.

Haploid kernels appear naturally, but are very rare. Haploid inducer lines must be used to increase the amount of haploids that appear on an ear of corn. Boote et al. (2016) reports a hybrid line crossed with an inducer line will produce roughly 10% haploids on the ear of corn. These haploid inducer lines are bred not just to induce haploids, but also to make them

distinguishable from the diploids. Roeber et al. (2005) reports that in the past, corn breeders have relied on the R1-nj gene expression to distinguish the two classes by visual inspection using the RWS/RWK-76 haploid inducer line. The R1-nj gene works by turning the endosperm and the embryo purple. If the inducer line is successful in merging with the embryo, the kernel will have a purple tinted embryo and endosperm, the marks of a diploid kernel. If the inducer line doesn't merge with the embryo, the endosperm will remain purple, but the embryo will be colorless, a signal of a haploid kernel (Boote et al. (2016)). Kernels pollinated by a third party are identifiable as well, since neither endosperm nor embryo will be purple. Figure 1.1 shows the difference between haploid and diploid kernels.

1.1.2 Haploid Maize Screening

There are visual cues in the full grown plant when the R1-nj gene is utilized. Generally the failed inductions will grow taller and healthier than the haploid (homozygous) kernel. Also the failed induction corn plant will have a purple base of the stem and root. This is beneficial, as it makes it clearly observable in the field which plants are haploid and which are diploid. This is a suboptimal solution however, because each diploid kernel planted is a waste of resources. Classification of haploids in the kernel stage is best.

Many researchers have endeavored to automate the haploid/diploid classification process using a variety of technologies. Each has its advantages and disadvantages. In this section we will look at a few possibilities. The natural extension of the visual inspection is to set up a hi-resolution camera and make the distinction using image processing. Techniques such as this are difficult for a variety of reasons. The geometry of the corn kernel is very non-uniform. The embryo will not always be in the same place or orientation in the kernel. Because of this, visual inspection systems must have a way of manipulating the kernels for optimal camera angles, thus reducing their effectiveness. Furthermore, if the genetic line contains such genes as the C1-I, the R1-nj gene will be completely repressed and visual inspection will be impossible (Melchinger et al. (2013)).

Most of the research in this field has been directed toward methods that: (a) better leverage present haploid markers (Jones et al. (2012); Boote et al. (2016); Fuente et al. (2017)), (b)

Figure 1.1: Visual haploid/diploid kernel discrimination is done by looking at both the embryo and the endosperm. Both kernels have purple endosperms, the right kernel has purple embryo and left does not. The left kernel is haploid, the right is diploid.

leverage the difference between haploid and diploid kernel that are independent of the expression of the marker (Smelser et al. (2015)), or (c) use different inducers entirely (Melchinger et al. (2013); Wang et al. (2016)).

Jones et al. (2012) found near-infrared spectroscopy could be used, but had trouble when multiple lines of corn were inspected simultaneously. With a two-step process, the analysis could be made reliable, in the 85% range. Boote et al. (2016) used fluorescence imaging, and Fuente et al. (2017) used multispectral imaging. It is not clear if these methods would work in the presence of the C1-I gene.

Smelser et al. (2015) used weight to make the distinction. They performed two experiments, one with two lines of corn, eleven haploids and eleven diploids (44 total kernels), and one with six lines, (132 total kernels). An ANOVA was run on both experimental sets. They found that haploid kernels weigh significantly less than diploid kernels when the analysis is kept to a single genetic line. Interaction between type of kernel, genetic line, and weight was significant. Because this method can be very fast, the authors theorize a technique like this can speed up the classification process compared to visual inspection alone.

Melchinger et al. (2013) explored the use of high oil inducers. They demonstrated a proof of concept for the use of oil content to distinguish between haploid and diploid kernels when a high oil inducer was available. This is notable because high oil inducers can produce classifiable haploids in lines where the R1-nj gene is suppressed. Nuclear magnetic resonance spectroscopy (NMR) is used to measure oil content in corn kernels. Wang et al. (2016) used NMR to separate haploid and diploid kernels by oil content ratio and weight. They built a high-throughput machine that could evaluate kernels at approximately four seconds per kernel. They achieved between 90% and 96% accuracy on six lines of corn. There is potential for terahertz radiation (THz) to be useful in distinguishing corn kernel haploids induced in this way. Liu (2017) demonstrated the use of THz scanning and chemometrics techniques for distinguishing between transgenic and non-transgenic corn oil.

1.1.3 Corn Kernel Sorting by THz and Machine Learning

The industry standard method for separating haploid from diploid kernels at present is by trained technician. This can be very expensive at the scale required. There is a need for automation in this industry. This work represents a proof of concept of a method to automate this task.

THz energy was the mode of kernel interrogation and data collection in this work. THz has great potential because Thz radiation can penetrate dielectric materials, exposing internal structure and composition to data collection and analysis. THz technology is relatively new and has exhibited great potential in a number of industries. A THz time-domain-spectroscopy (TDS) system was used to collect time-domain waveforms.

Preliminary data such as the refractive index of the corn kernel and an approximate scan volume were calculated. Each kernel was then raster scanned, approximately 51 worthwhile waveforms were collected per kernel. Each waveform is transformed to the frequency domain by Fourier transform, and each frequency is treated like a separate dimension in the data. This data set is both large and multivariate, requiring a machine learning (ML) model for analysis. A probabilistic neural network (PNN) was used to classify each waveform, and by proxy each kernel, as either haploid or diploid. In this way the classification process can be automated.

1.2 THz Technology

The terahertz (THz) band is part of the electromagnetic spectrum between 0.1 and 10 THz. It lies between the microwave and infrared regimes and as such, it shares properties with both. Like microwave, THz can penetrate most materials, provided they have low electric conductivity and are non-polar. Like infrared, the THz band contains absorbance lines for many materials in the frequency domain. These two factors lead to the strength of THz as both a time-domain and frequency-domain tool. As non-ionizing radiation, THz does not impose a health risk like x-ray. It can be used for spectroscopy with applications in chemistry, astronomy, and medical industries. It can also be used as an imaging tool, with applications in thickness measurements and non-destructive evaluation. Mittleman (2018) is a good review of the state of the art of THz imaging.

The terahertz gap has been a subject of intense research in the past half century. It has been a difficult regime to explore. Under the term "sub-millimeter waves", it has been investigated from the microwave electronic perspective. Microwave electronics have been developed that push the envelope into frequencies as high as 100 GHz with the use of Gunn and IMPATT diodes. Meanwhile the THz band lies on the low end of the infrared regime, under the term "extremely far infrared". Some infrared spectrometers have reached down to the tens of THz in frequency. Development in the regime was limited since there were no known applications for such EM frequencies. In the field of nondestructive evaluation (NDE) researchers are still looking for the "killer application" that will bring THz-NDE into the limelight. In recent years, however, the topic has seen a resurgence as more methods of THz generation have developed.

The study of astronomy drove most of the development of THz technology in the early years. THz frequency light corresponds to the low tens to hundreds of Kelvin in black body temperature. Cooling interstellar dust clouds correspond to this temperature, making THz a tool for probing the galaxy (Siegel (2002)). Spectroscopic interests drove development in THz science as well. Absorption in this regime corresponds with mass dependent rotation of gasses (De Lucia (2003)). Collective vibrational and torsional modes in condensed media can be probed. Shen et al. (2005) could distinguish explosives using their spectral fingerprint in

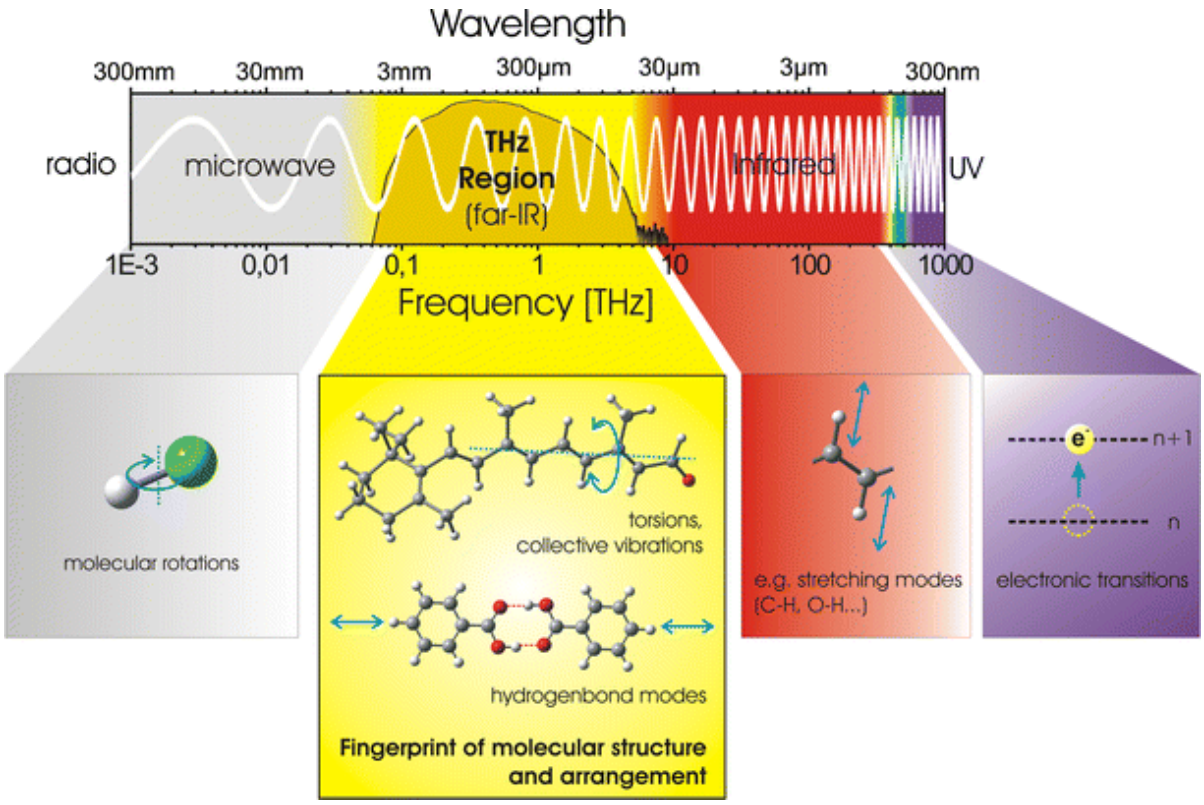


Figure 1.2: Electromagnetic spectrum with THz band highlighted (as seen in Walther et al. (2010)).

the THz regime. What is helpful for spectroscopy is detrimental for applications such as radar, imaging, and communications. The atmosphere readily attenuates THz radiation, especially water vapor. Without high power sources or dry conditions, THz has a small propagation distance in the atmosphere. THz sensitivity to water has been exploited in a few studies. Banerjee et al. (2008) explored the use of THz radiation for determining the water content of paper. Castro-Camus et al. (2013) used THz spectroscopy data to observe plant hydration under water restriction.

The proliferation of ultrafast femtosecond lasers opened the door to many new technologies related to THz. New sources and detectors utilizing these lasers have been developed (Mittleman (2003)). Continuous wave (CW) systems and time domain spectroscopy (TDS) systems, by way of photoconductive antennas, will be discussed in the following paragraphs.

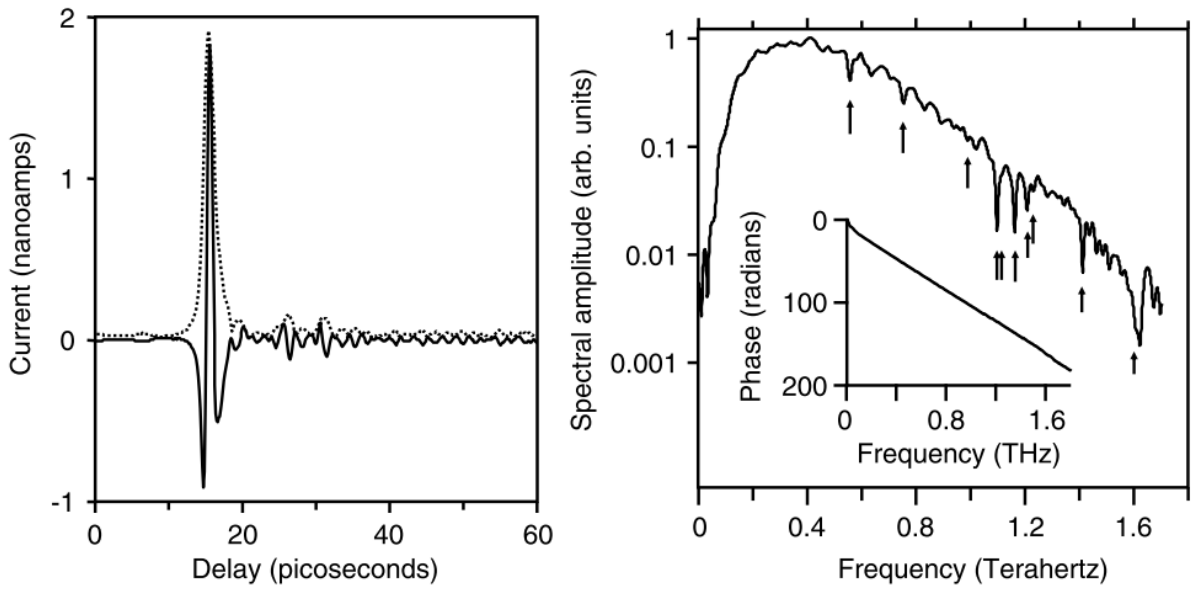


Figure 1.3: Water readily attenuates THz frequency radiation. As seen in Chan et al. (2007).

Table 1.1: Pulsed THz and continuous wave system comparison, (adapted from Zhang (2010)).

	cw-THz wave imaging	Pulsed THz wave imaging
Cost	$\$50,000$ - $\$150,000$	$\$200,000$ - $\$1,000,000$
Weight	3 kg	10 kg
Speed	100,000 points/s	$< 4,000$ points/s
Spectral Information	No	Yes
Depth Information	No	Yes
Refractive Index	No	Yes

CW systems boast tremendous capability in spectroscopic applications. One method of producing THz energy for such a system is called "photomixing" or "heterodyning". In a photomixing source, two high frequency infrared lasers are used with slightly different operating frequencies. Mixing these lasers produces a beat frequency in the THz regime. Directing these lasers at a photomixer will result in an electromagnetic beam produced at that beat frequency. Low temperature grown gallium aresenide (LT-GaAs) photoconductive antennas are used to emit and detect electromagnetic waves in this range (Verghese (1997)). This method provides a small bandwidth continuous THz wave that is tunable according to the frequencies of the

pump lasers. Ferguson and Zhang (2002) and Consolino et al. (2017) provide good reviews of modern CW sources and their applications to molecular spectroscopy.

The system used in this work is of the THz-TDS type. The main features haven't changed appreciably since its introduction in 1989 (Smith et al. (1988); Fattinger and Grischkowsky (1989)). They include a femtosecond laser, optical delay line, THz emitter and detector. As opposed to CW systems, TDS systems are characterized by high bandwidth pulses and fast capture times. Figure 1.4 details the structure of a TDS system.

The lasers generally have pulse lengths in the 10 femtosecond range, while repetition rates can be in the 100 MHz range. The laser beam is split at a beam splitter, one pulse directed toward the transmitter and one directed toward the receiver. A biased photoconductive antenna is used to generate the pulse. This is usually a dipole antenna printed on an LT-GaAs semiconductor, with additional substrate lenses to collimate the beam. Further lenses are required to focus the beam. An unbiased photoconductive antenna is used to sample the electric field at the receiver. Figure 1.5 shows how the TDS may collect data.

With every pulse of the femtosecond laser, one pulse of THz energy is produced and one data point of the waveform is collected. The optical delay line cycles back and forth changing the path length of the laser pulse on its way to the receiver. This changes the timing of its data collection. In this way a number of points can be sampled on the waveform (Chan et al. (2007)). The timing is critical, and very precisely tuned.

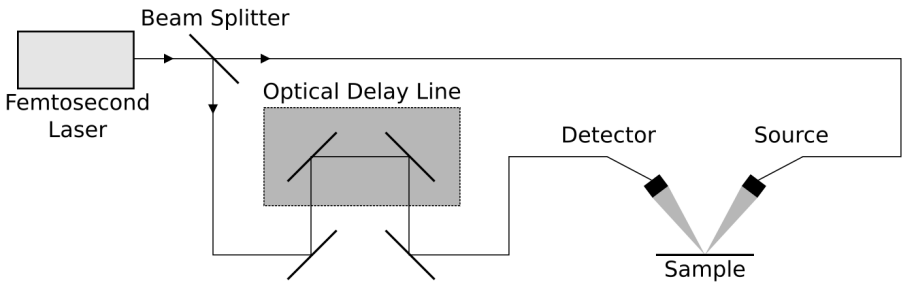


Figure 1.4: Schematic of a THz-TDS system.

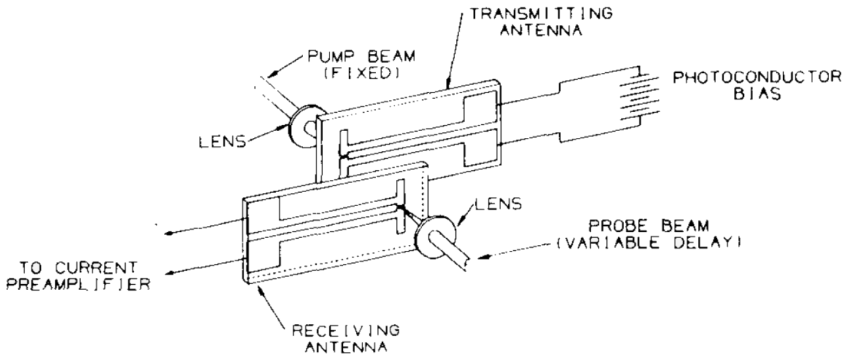


Figure 1.5: Photoconductve antennas (Smith et al. (1988)).

Sampling the THz pulse produces amplitude and phase information. This data is useful in the time-domain for the measurement of spatial properties such as thickness, distance, and scattering from discontinuities. The data can be readily transformed to the frequency domain via Fourier transform in order to analyze absorption and dispersion. More information about CW systems, TDS systems, and THz science in general can be found in Lee (2009).

The first reported use of TDS for imaging was by Hu and Nuss (1995). Since then interest in imaging applications has exploded. The capability of THz-TDS was demonstrated by imaging integrated circuit packages as well as biological materials such as a leaf. The plastic packaging on the integrated circuit is transparent to THz radiation, making it easy to see the embedded metal components. The leaf demonstrated the sensitivity and absorption of THz waves interacting with water. Two images were taken of the same leaf, one while it was fresh, and another after 48 hours. The first showed large absorption due to water present in the leaf, the second showed much less, except in the veins of the leaf; showing how the water had evaporated.

The unit of data analysis in THz-TDS is the A-scan, a time resolved estimation of the pulse's electric field. Raster scanning a sample produces a three dimensional data set that can be explored through B-scans and C-scans. C-scans are a slice in time, the resulting image is an amplitude plot across both spatial dimensions. B-scans are a slice in a spatial dimension, with an image of the other spatial dimension and time. Fourier transformation can be done as well, turning the time axis to a frequency axis. Figure 1.6 shows what this data set looks like.

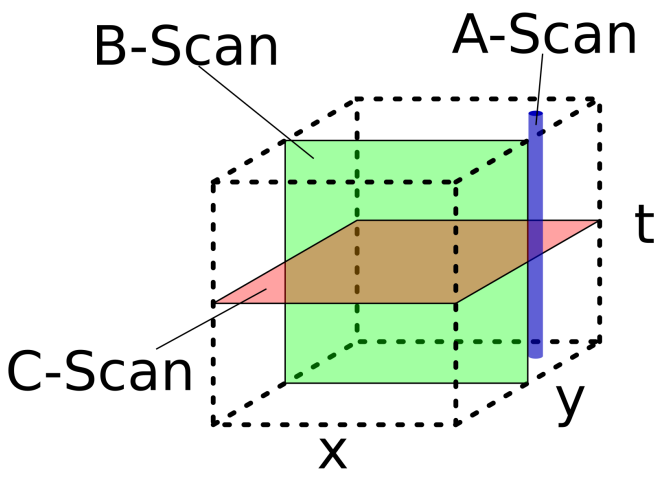


Figure 1.6: THz-TDS data structure can be visualized three ways, with A-scans, B-scans, and C-scans.

1.3 THz Nondestructive Evaluation

THz nondestructive evaluation (THz-NDE) is a non-contact method requiring no couplant, which is a great benefit compared to other NDE methods. Changes in wave impedance is the source of reflections in any wavelike phenomenon. Changes in electromagnetic (EM) wave impedance across interfaces is the source of variability in reflective amplitude, and thus contrast in a THz scan. EM wave impedance is written as:

$$Z = \sqrt{\frac{i\omega\mu}{\sigma + i\omega\epsilon}} \tag{1.1}$$

where ω is the angular frequency, μ is the magnetic permeability, ϵ is the electric permittivity, and σ is the electric conductivity of the medium. The impedance of a void is much lower than that of a solid in terms of electromagnetic waves, thus they are readily detected. An ideal dielectric has an EM wave impedance given by:

$$Z = \sqrt{\frac{\mu}{\epsilon}} \tag{1.2}$$

as the conductivity decreases to zero. Most dielectrics have a relative magnetic permeability close to one, therefore contrast arises primarily from changes in electric permittivity. NDE of non-metallic materials is the ideal use case of THz radiation.

THz-NDE relates to microwave NDE in many ways. Both methods are non-contact and readily pass through air. Due to its high frequency, the skin depth in conductive media is smaller than in microwave, but the resolution is finer. The methods most common in THz-NDE use optical components instead of waveguides such as in microwave NDE. Microwave technology is more mature than THz technology. This is evident in the amount of solid state microwave electronics there are. Further information regarding the physics, limitations, and NDE modalities of microwave NDE can be found in a review paper by Rollwitz (1989).

While the space shuttles were active THz-NDE was applied to the inspection of foam insulation. The Columbia disaster was caused by defects in the foam on the external fuel tank. Zimdars et al. (2005) showed that THz pulses scatter from voids and reflect from disbonds within the foam, with a total penetration depth of 20 cm at 300 GHz.

The aerospace industry has seen a huge increase in the use of composites. Modern aircraft like the Boeing 787 can have as much as 50% composite by weight (Boeing (2018)). Carbon fiber composites have high conductivity along the fibers, limiting the capability of THz inspection. Conversely, glass fiber reinforced composites (GFRC) are readily inspectable via THz.

Stoik et al. (2008) examined aircraft composites with defects such as voids, delaminations, mechanical damage, and heat damage. Hsu et al. (2011) demonstrate the ability of THz inspection to detect defects behind other defects, a limitation of ultrasonic inspection due to the "shadow effect". Lopato and Chady (2013) show how THz-NDE can be used to find damage caused by mechanical impact in basalt fiber reinforced composites.

Chiou et al. (2012) applied THz-NDE to the inspection of ultra high molecular weight polyethylene (UHMWPE) armor plates for manufactured defects. Palka et al. (2015) investigated the use of THz to inspect the damage done to UHMWPE armor plates by firearms. They found they could identify stages of the interaction between the bullet and the armor, such as a shearing and total destruction of the material for a few millimeters, followed by a blunting of the bullet and a stretching of the composite to "catch" the bullet. They found a THz-TDS raster scanning method adequate for the analysis.

Yakovlev et al. (2016) demonstrate how THz-TDS can be used in manufacturing GFRCs to observe the curing process. They calculated the material parameters of the epoxy alone at

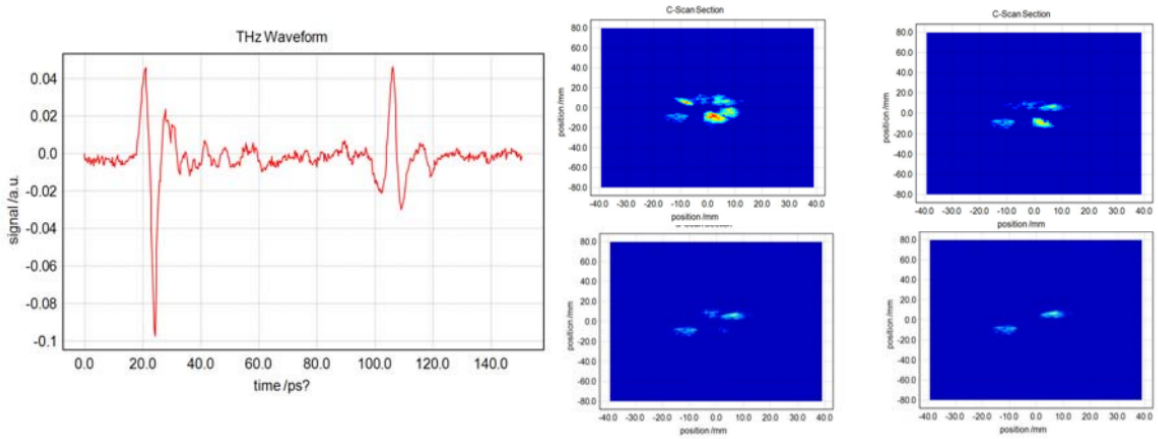


Figure 1.7: Time-domain TDS waveform of glass composite laminate (a) and impact delamination images using different time gates (from left: (b) 4369 ps, (c) 4769 ps, and (d) 5669 ps) (Hsu et al. (2011))

various stages of curing. They found THz-TDS data is adequate for calculating these material parameters. The material parameters match well with the curing process, allowing THz-TDS scanning to monitor the curing process.

Radomes are protective covers for sensing and radar equipment on airplanes. Usually made from composites, they are specifically designed to protect the electronics without interfering with their function. However, they must be transparent at the operating frequencies of the antenna inside. Modern fighter aircraft are designed to have as low radar cross-section as possible; something a completely electromagnetically transparent material will not provide. New developments are being made in radomes that are transparent at the operating frequency of the internal antennas, but reduce the radar cross-section outside the antenna frequency range. Panwar (2018) reviews many techniques for evaluating the electromagnetic properties, as well as the non-destructive evaluation, of radomes. Many of these techniques fall within the realm of terahertz technology.

Ample work has been done applying THz-NDE to the inspection of coatings. Anastasi and Madaras (2006) performed surface roughness characterization with and without multiple layers of paint with THz. Marine and automotive paint have been successfully inspected using

THz, as well as sustained release coatings on pharmaceutical tablets (Zhong (2018)). Catapano et al. (2017) explored THz inspection of icephobic coatings for aircraft. Fukuchi et al. (2016) applied THz-NDE to the inspection of thermal barrier coatings (TBC). TBCs are ceramic coatings on metal substrate designed to withstand extremely high temperatures and protect the metal components they coat. They are typically found in high temperature environments such as inside gas turbine engines. Types of damage that may occur on TBCs include topcoat thinning, topcoat delamination, thermally grown oxide layer between topcoat and substrate, and surface cracking. TBC thickness measurements have been demonstrated using THz-TDS. Chen et al. (2010) reports thermally grown oxide layer defects on TBCs can be monitored using THz time-domain reflectometry.

1.3.1 Material Characterization

Material characterization is a frequent topic of NDE and has been given much attention in the THz space. The work was pioneered by Duvillaret et al. (1996). In their work, the inverse problem of finding the refractive index from the data could be solved reliably. Equation 1.3 shows the form of the complex refractive index.

$$\hat{n}_a = n_a - ik_a \tag{1.3}$$

They showed that the complex refractive index could be found for both optically thick samples, where the echoes of the THz pulse are well separated, and optically thin samples where THz echo pulses overlap.

In a later work, Duvillaret et al. (1999) demonstrated that material thickness as well as index of refraction could be determined. Others have expanded on the work including Dorney et al. (2001), Pupeza et al. (2007), and Hejase (2012).

One interesting study by Busch et al. (2014) examined the optical properties of materials used for fused deposition modeling 3D printing. They tested seven materials for refractive index and absorption coefficient. Table 1.2 shows the results. High Density Polyethylene (HDPE) and Polypropylene (PP) have absorption coefficients near zero, making them good candidates for THz quasi-optical components. However, these two materials are very hard to print reliably.

A good compromise of reliability and absorption coefficient is Polystyrene. They produce lenses for the THz range and demonstrate their effectiveness.

1.4 Other THz Applications

THz technologies have been used in the food and agriculture industries for a variety of purposes. Quality control is a big driver. Mathanker et al. (2013), Wang et al. (2017), and Wang et al. (2018) reviewed THz applications in the food and agriculture industries.

Ge et al. (2014) showed THz-TDS could reliably be used to identify wheat quality, when combined with chemometric tools. The study involved four quality levels: normal, moldy, worm-eaten, and sprouting wheat grain. Spectra in the 0.2 - 1.6 THz range, processed with a combination of PCA and support vector machine (SVM), can be used with as high as 96.7% prediction accuracy.

Xu et al. (2015) used THz-TDS and chemometrics to differentiate transgenic rice from its parent rice. The Hun Hui 1 rice is genetically modified to produce the protein Cry1Ab as a way to deter pests. The parent line, Ming Hui 63, has no such capability. Genetically modified organisms (GMOs) are heavily regulated in many countries. Xu et al. (2015) showed that THz-TDS data, paired with chemometrics techniques, could achieve as high as 85.3% accuracy distinguishing GMO rice from its parent line. Pretreatment of the spectra and discriminant analysis were successful methods. Methods such as principal component analysis (PCA) and partial least squares discriminant analysis (PLS-DA) performed poorly.

Gente et al. (2016) used THz-TDS to differentiate proper and defective beet seeds. In the field today x-ray imaging is used to find a weight threshold for each batch of beet seeds. Gente et al. (2016) found there is a large difference between the signals of proper and defective beet seeds. They theorize that the difference in signal is due to a difference in water content. Figure 1.8 shows the distinction that can be made.

THz has been applied to the corn industry as well. Sun et al. (2010) used Thz-TDS to classify two types of corn by examining estimates of absorption. In their work the embryos of the corn kernels were ground up prior to data collection. Lian et al. (2017) used THz-TDS to classify four types of transgenic corn, grinding and compressing the entire kernel into a tablet.

Table 1.2: Printing parameters of the investigated materials and the material parameters at 500 GHz (Busch et al. (2014)).

Material	Printing Temperature [°C]	Sample Thickness [mm]	Refractive Index n @ 500 GHz	Absorption Coefficient [cm^{-1}] @ 500 GHz
ABS	250	1	1.57	5
PLA	220	1	1.89	11
Nylon	255	1	1.72	9
Bendlay	250	5	1.532	1.8
Polystyrene	240	5	1.561	0.5
HDPE	230	5	1.532	0
PP	230	5	1.495	0

Hilscher et al. (2018) is a patent on seed classification in any way using THz-TDS data. The patent includes corn seed classification and any data processing technique, including machine learning. Haploid and diploid classification is not mentioned explicitly in the patent.

Other applications of THz science outside of NDE include the medical industry (Taylor et al. (2011); Pickwell et al. (2004)); communications (Suen (2016)), astronomy (Marrone et al. (2004)), security (Federici et al. (2005)), and art (Fukunaga and Picollo (2010); Catapano and Soldovieri (2017); Niijima et al. (2018)). The reach of THz technology expands every year.

1.5 Machine learning

Machine learning is a way to investigate data using computer algorithms that can improve in performance without being explicitly programmed. As a form of artificial intelligence, it borrows heavily from such disciplines as statistics and optimization. Large or complex data sets can be analyzed for purposes such as regression, classification, or clustering. Examples include: facial recognition, spam email filtering, optical character recognition, etc. Machine learning is used heavily in the field of chemometrics when experimental data is large and complicated, and an inversion must be performed to determine experimental conditions.

There are two main types of machine learning in the field: supervised and unsupervised. Unsupervised learning is looking for trends or patterns in unknown data. These methods are exploratory in nature. Data mining is a common synonym for this type of machine learn-

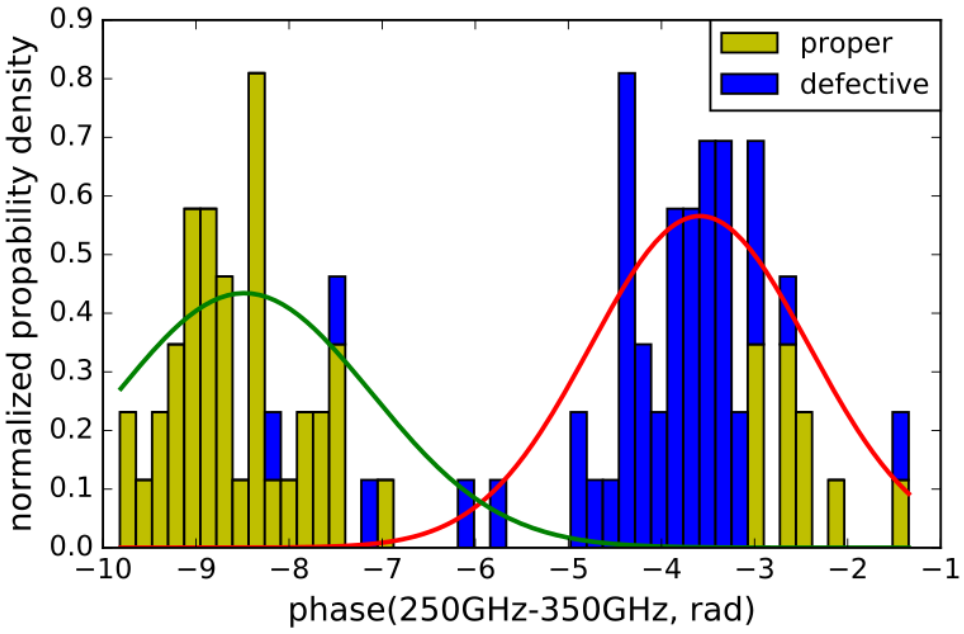


Figure 1.8: Defective and proper beet seeds can be distinguished using THz-TDS (Gente et al. (2016)).

ing. While supervised machine learning might be looking for a specific outcome or answer, unsupervised learning is not guided, and in some cases may only tell of groupings in data. Dimensionality reduction methods such as principal component analysis (PCA) fit under this term as well. PCA uses an orthogonal transformation to create a set of orthogonal variables from a set of possibly correlated variables. The first component represents a linear combination of the original variables with the highest possible variance. Each subsequent component has the highest possible variance with the added condition that it must be orthogonal to all previous components. With this method, a few variables can be used that correspond to a majority of the variance in a multivariate data set.

Supervised machine learning techniques can be used on data where there is a sample for which the pattern or data labels are known. The known data are used to train the machine learning model so it will be effective on new and unknown data. In the case of the spam email filter, a set of known spam can be used to train the model to recognize spam in an unknown batch of emails. A probabilistic neural network is of this type, and is the primary model used in this work.

Each model has parameters that define its behavior. These parameters must be optimized to improve the performance of the model on unknown (or test) data. A danger in the training stage is over-fitting a model. In this scenario, a model can be very well trained to work with the training data, perhaps 100% effective, but dismally effective when new data is introduced. Appropriate cross-validation (CV) can reduce this problem.

CV is the system for evaluating a supervised machine learning model using known data. Two CV methods were used in this work. The first, leave-one-out (LOO) works by setting aside a single data point, training with the others, and testing the model using the data point set aside previously. This method is very simple, but expensive in practice, growing more costly as the size of the training set increases. It is the same as the statistical resampling tool called the jackknife. The second model used is called K-folds, where the training set is divided into K groups. It follows the same steps as LOO, except it is active on the set of folds instead of the raw data. Each group is set aside in turn for testing, while all others are used for training. With these two methods, the performance of a machine learning model when introduced to new data can be evaluated using only information previously known (Varmuza and Filzmoser (2009)).

1.5.1 Probabilistic Neural Network

The probabilistic neural network (PNN) was introduced in Specht (1990). It is closely related to what Mitchell calls "instance based learning" using gaussian radial basis functions (Mitchell (1997)). It is a modified artificial neural network that can map any input to any number of output discrete classifications. The underlying equation uses a probability density function developed by Parzen to calculate decision boundaries. These boundaries can be linear or non-linear, dependent on the data set. A PNN stands at the edge between more traditional artificial neural networks called "eager learners", with extensive training stages up front, and methods called "lazy", where training doesn't happen until a new data point is presented.

Due to its simplicity and ease of use, the PNN has been applied to many areas with great success. Steenhoek (1999) used a PNN to evaluate corn kernel damage. Images of corn kernels

with blue-eye mold damage, germ damage, and no damage, were used to train the model. Accuracy for the blue-eye mold damaged category was 78%, while the germ-damaged and no damage categories were 94% and 93%, respectively.

Garoudja et al. (2017) used a pair of PNNs to recognize and classify faults in an array of photovoltaic solar cells. They compared traditional back-propagation artificial neural networks to the PNN and found the PNN was more accurate and resistant to noisy data. They found the PNN was 82.34% accurate in classifying a faulty state in the array, and 98.19% accurate in classifying the type of fault that had occurred.

Other applications include Bastke (2009) who used a PNN to detect anomalies in computer networks, Vinodhini and Chandrasekaran (2016) used a PNN for sentiment classification of online reviews, and Jebarani and Kamalaharidharini (2017) used a PNN for human face recognition, among others.

1.6 Thesis Structure

The rest of this thesis contains information on the methods, results, and conclusions of this work. Sections 2.1 and 2.2 include information about the experimental capabilities available, and the setup of the corn kernels for data collection. Sections 2.3 to 2.5 are about the steps taken in preprocessing the data. Sections 2.6 to 2.10 explore how a probabilistic neural network works, how it was implemented, and its behavior under certain scenarios. Chapter 3 delves into the results, including a preliminary study of the electromagnetic properties of a corn kernel, some estimates of performance and stability of the model, and also some discussions of its implications. Finally, chapter 4 includes some final thoughts and take-aways of this work.

CHAPTER 2. METHODS

In the following sections, the methods used in this work are described including: how Thz corn kernel data were collected using a THz-TDS system, how the data were processed to organize and reduce the data volume, and finally how the probabilistic neural network was trained and applied, including validation of model behavior.

2.1 Experimental Capabilities

The THz-TDS system used in this work for data collection is called a "TPI Imaga", by Teraview. It uses a mode-locked 100 femtosecond Ti:Sapphire laser to drive a photoconductive antenna using a lock-in amplifier. Detection is done using a similar photoconductive antenna. The beam is highly focused at 50 mm focal length with a beamwidth at full width half max of 0.8 mm. It is oriented for reflection mode at 17° from the normal; as can be seen in figure 2.1. The transmitter and receiver are mounted on a 3D gantry capable of 2D translation and raster scanning. An ad-hoc plastic tent was used to isolate the air around the beam. Dry air was pumped into the tent to limit the interference of water vapor. Figure 2.2 shows the THz-TDS system in the THz lab at the Center for Nondestructive Evaluation.

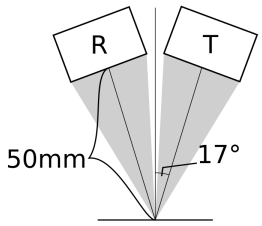


Figure 2.1: The scanner is set in reflection mode with 50 mm of focal length at 17° from normal.

A time gate must be chosen that defines the exact time window, relative to the laser pulse activation, within which the waveform will be recorded. A priori knowledge of the scan target such as surface height, surface variability, and refractive index are used to define the start and end of the gate such that all the data is collected properly. After the time gate has been set, an A-scan of a metal plate perpendicular to the reflection plane was used as a reference. This reference was taken with the metal plate in focus. The peak of the reference waveform represents the max amplitude that can be expected. The arrival time can be used to reference when a signal is in focus, provided the beam path passes through only air.

These reference waveforms also serve to relate experiments across time. The THz-TDS hardware performance may change day to day, varying as much as 10% under normal operating conditions. References from two experimental sessions can be used to normalize the collected waveforms and compare directly.

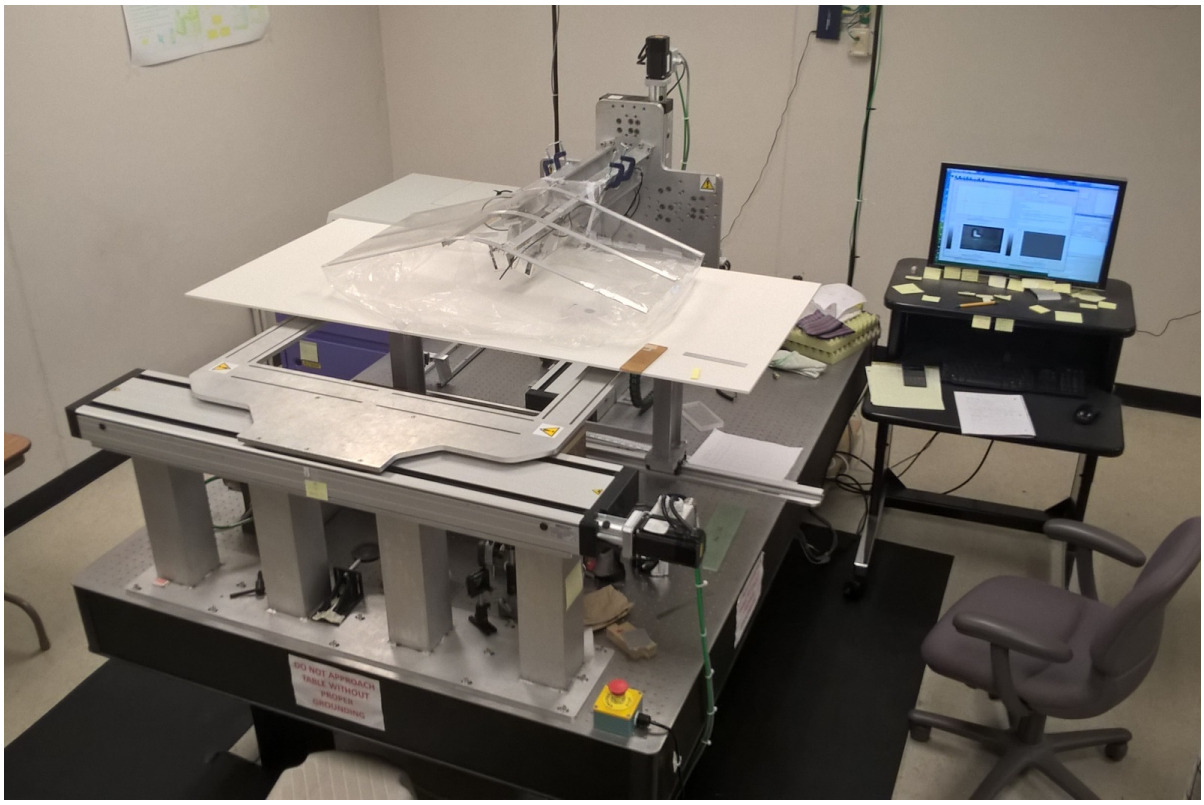


Figure 2.2: THz-TDS system in the THz lab.

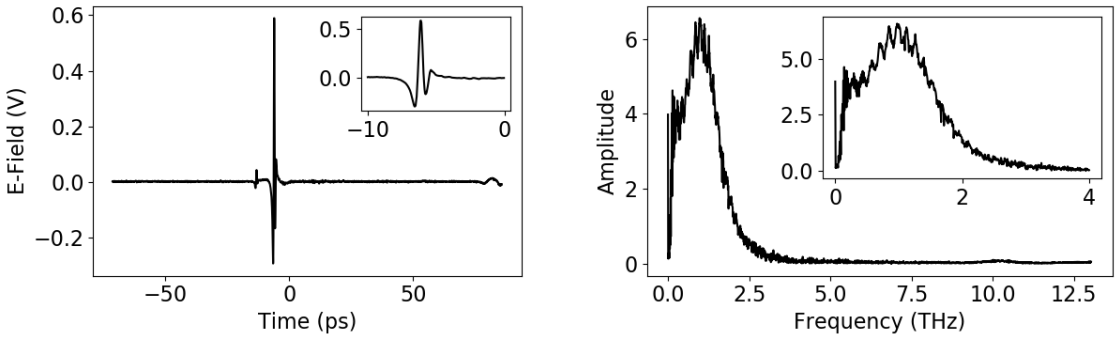


Figure 2.3: Reference waveform in the time domain (left) and spectrum in the frequency domain (right). Inset images are simply cropped and magnified from the main images.

2.2 Sample Corn Kernels

The sample set consists of 91 hand picked corn kernels. They were carefully selected with two goals in mind, first, they had to be readily classifiable haploid or diploid after visual inspection, and second, they had to have a relatively flat front face to lay perpendicular to the scan axis. This was to make the most favorable conditions possible for scanning the embryo. It is understood that in practice these conditions will not be constantly met. The kernels were laid out in a grid, sufficiently spaced so that the signal would go to zero between kernels. This helped in image segmentation and data labeling later. In a future implementation this orderly layout can be mechanically arranged on a conveyor belt, for example. Figure 2.4 shows the corn kernels used in the analysis.

The corn kernels came from five sources. Table 2.1 details the collection of kernels. Five different F1 (hybrid) corn lines were crossed with the RWS/RWK-76 haploid inducer line (Roeber et al. (2005)).

They were separated into haploid and diploid according to visual markers provided by the R1-nj gene. These data form the labels that will be used for training the PNN. The data set is built by collecting a time-domain waveform (A-scan) from each point in a 2D image with resolution of 0.5 mm x 0.5 mm. Figure 2.7 shows an image where each pixel represents the

Figure 2.4: Corn kernel sample used in the analysis.

Table 2.1: Corn line designations, classification, and count in the sample.

Row	**Designation**	**# Haploid**	**# Diploid**
1	GF6/GF3	9	8
2	GF1/GF3	9	9
3	GF4/GF3	10	9
4	GF1/GF4	10	9
5	GF2/GF1	9	9

max amplitude of an A-scan taken at that location. Each kernel has on average 251 A-scans associated with it. Table 2.2 shows the details of the data set parameters.

The time gate for capturing the waveforms was very large compared to the size of the waveform needed to perform the analysis. This was set in order to accomodate the varying geometry of the corn kernels, and ease the post processing by including reference points. The gate had to be wide enough that it could capture the signal from the tallest corn kernel, but also the signal from the baseplate when the scanner was between kernels. This led to a choice of time gate of 157 picoseconds. This time gate was sampled 4096 times, leading to a sampling time step of 38.33 femtoseconds, or a sampling frequency of about 26 THz.

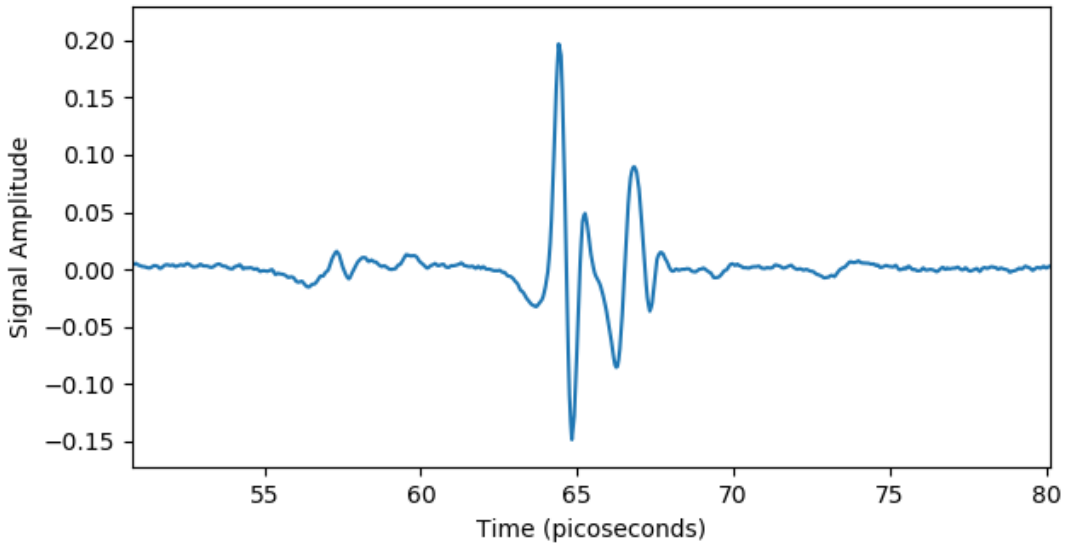


Figure 2.5: A waveform taken on the surface of a kernel. Because of the geometry and random nature of the corn kernel, the waveform can be misshapen.

Table 2.2: Data collection parameters.

Parameter	**Value**
Scan Resolution	0.5 mm x 0.5 mm
Number of Scans	4
Scan Height	4, 6, 7, 8 mm from baseplate
Total A-Scans on Kernel Faces	19,533
A-scans Per Kernel	215
Time Resolution	38.3 femtoseconds
Total Kernels, Diploid, Haploid	91, 44, 47

2.3 Focus Compensation

The complicated shape of the corn kernel has imposed significant technical difficulty in this work. The kernel shape and thickness/height can vary drastically and randomly from kernel to kernel. The tallest kernel has a height of 7.49 mm, while the smallest kernel is only 3.87 mm tall. Figure 2.6 shows the variance in the kernel max height. As an aside, it appears a distinction can be made between haploid and diploid kernels using only these data, but the overlap would be severe, leading to a high error rate. Such variance in height has caused large

fluctuations in the signal because the beam is tightly focused. To mitigate this problem, four scans were performed with focus set at 4 mm, 6 mm, 7 mm, and 8 mm height measured from the sample plate. This produced four data sets. The four were reduced to one by comparing peak-to-peak amplitude of the four choices at each pixel and saving the waveform with highest peak-to-peak amplitude. Figure 2.7 shows this process.

2.4 Image Segmentation

Image segmentation is the image processing term for identifying and separating regions of an image. Common algorithms range from simple thresholding and edge detection to complex model based segmentation.

A time-of-flight image was created from the data set, where each pixel represents the time difference between the start of data recording and when the waveform peaks. This creates an image that correlates with the height of the kernel, tall kernels represent a low time-of-flight, and short kernels represent high time-of-flight. An image segmentation routine was used to separate the kernel boundaries on the time-of-flight image and the corresponding waveform data. Some kernels were too close to be separated by contrast in the time-of-flight images alone. This problem was solved by setting a threshold, applying a distance transform (Jones et al. (2017)), and applying a watershed image segmentation algorithm (van der Walt et al. (2014); Neubert and Protzel (2014)).

The threshold separated the kernel face pixels from the baseplate pixels and made a binary image. The distance transform gave a value to each pixel equal to its distance from the closest edge. This created a high point at the center of a kernel, going to zero at the edges. After the distance transform, pixels at the intersection of two kernels will form a valley between kernel centers. The negative of this image formed the starting point for the watershed segmentation algorithm. The watershed transformation is a type of region-growing image segmentation that uses the gradient of the image to define catch-basins used for labeling pixels. A watershed transformation takes a series of labeled pixels and treats them like “fountains”, labeling surrounding pixels as the “water” rises. With one fountain at the base of each kernel, the surrounding pixels will be labeled the same, but will stop the process when the “water” touches a pixel claimed

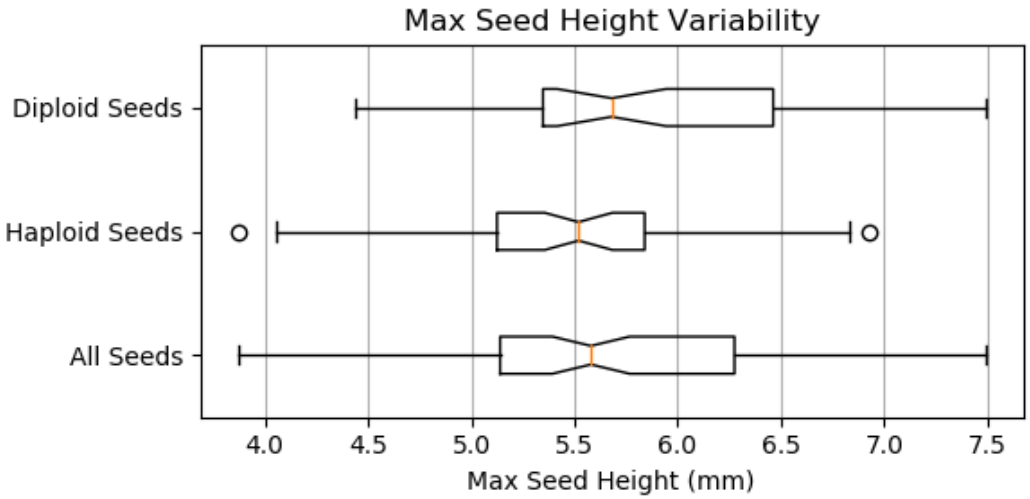


Figure 2.6: Box plot showing the variability of the height of the kernels.

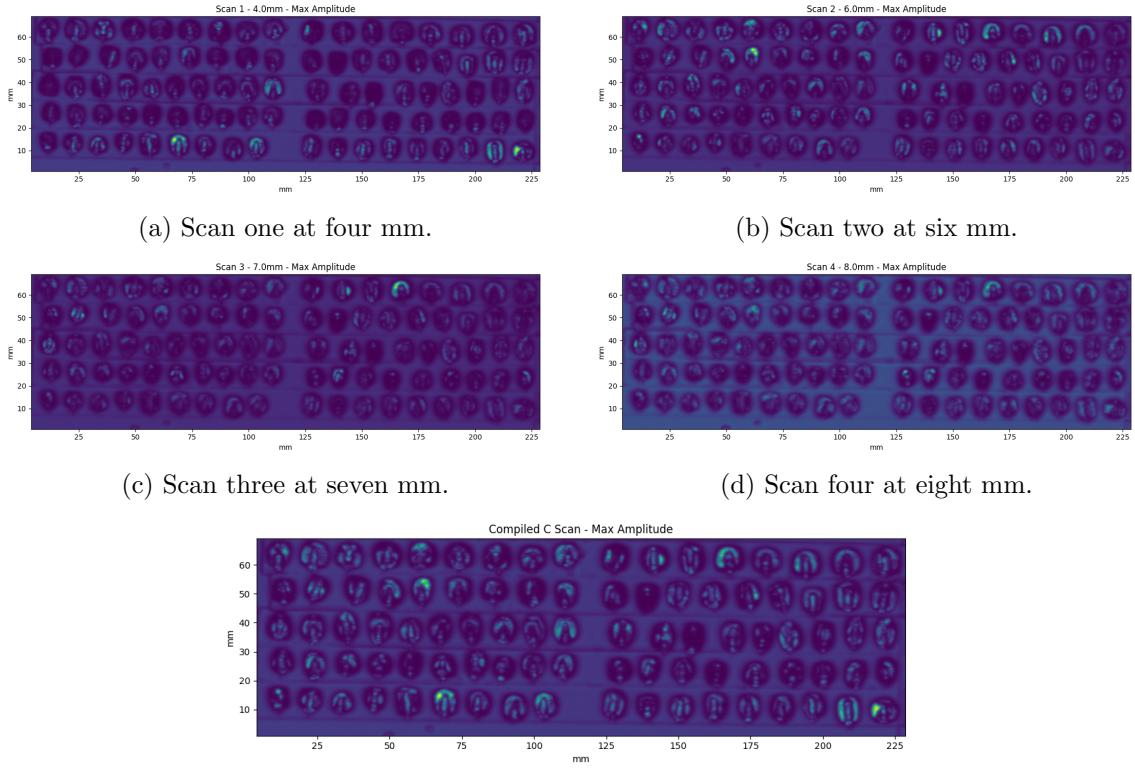
by another label. In this way all 91 kernels and corresponding waveform data can be labeled and segmented. Figure 2.9 shows the results of this image segmentation.

2.5 Data Reduction and Transformation

This stage of work seeks to distinguish haploid from diploid kernels using the spectra from each A-scan. A fast Fourier transform (fft) function in the python package Scipy (Jones et al. (2017)) was used to produce these spectra. The magnitude of the spectra was used in this work. Equation 2.5 shows the equation for an fft, and equation 2.5 shows the equation for the ifft, the inverse fast Fourier transform.

$$Y_k = \sum_{n=0}^{N-1} x_n exp\left[-2\pi i \frac{kn}{N}\right] \tag{2.1}$$

$$x_n = \frac{1}{N} \sum_{k=0}^{N-1} Y_k exp\left[2\pi i \frac{kn}{N}\right] \tag{2.2}$$



(a) Scan one at four mm.

(b) Scan two at six mm.

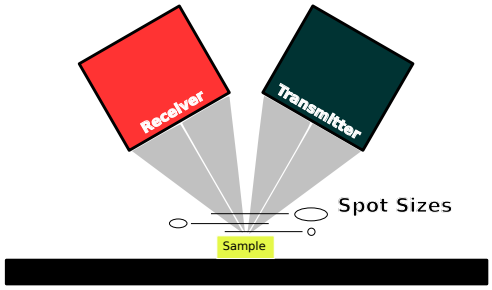
(c) Scan three at seven mm.

(d) Scan four at eight mm.

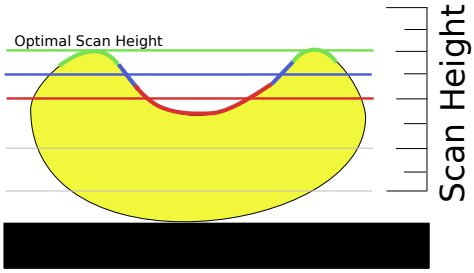
(e) Compiled data set.

Figure 2.7: Four scans provide the basis for one compiled data set. The compiled data set was derived from the four choices based on max amplitude in the time domain.

Since the THz system can only produce spectra up to 4.0 THz reliably, all numerical spectral data were truncated at 4 THz. With further testing, it was discovered that lowering the upper frequency limit below 4.0 THz produced better results; more on this in chapter 3. The data set was further reduced by making the assumption that the important information will be found in the embryo. The data set was reduced to only those pixels approximately related to the embryo, as can be seen in figure 2.10. In that figure, the black pixels represent spectra identified in preprocessing as associated with a kernel, and the red pixels represent spectra approximately on the embryo. This reduces the number of spectra on average from 215 to a more tractable 51 per kernel.



(a) Focal zone has the smallest spot size.



(b) Different scan heights allow for different kernel surfaces to be in focus.

Figure 2.8: Multiple scans were used to ensure each pixel had a scan that was in focus.

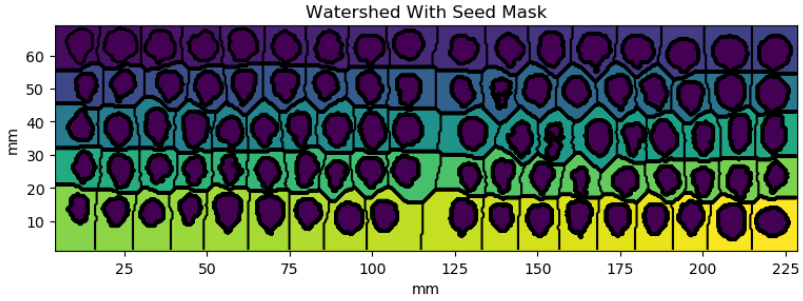


Figure 2.9: Each color marks a different kernel label; this illustrates how each data point is assigned to a kernel.

2.6 PNN Theory

Parzen (1962) showed that a class of PDF estimators approach the parent density function so long as it is continuous. This estimation of the probability allows the use of the Bayes rule, and thereby create a Bayes learner. This work is a binary decision ($d(X)$) made by way of PNN, having the rule

$$d(X) = \theta_A \text{ if } h_A l_A f_A(X) > h_B l_B f_B(X) \tag{2.3}$$

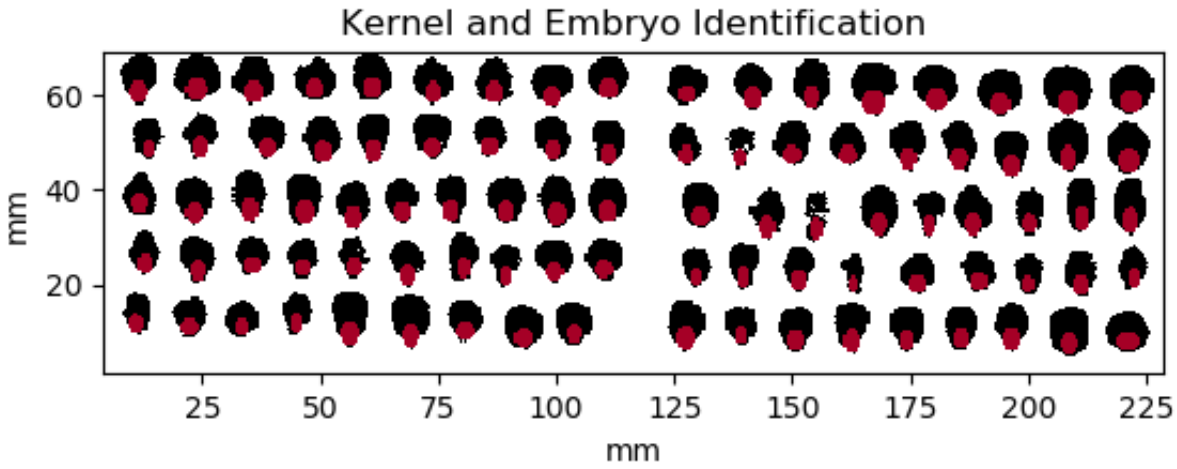


Figure 2.10: The red areas represent the embryos and the data in these areas were used in the model.

$$d(X) = \theta_B \text{ if } h_A l_A f_A(X) < h_B l_B f_B(X) \tag{2.4}$$

where θ_A and θ_B are the output classification decision of either class A or class B; $f_A(X)$ and $f_B(X)$ are the probability density functions of classes A and B, respectively; l_A and l_B are the loss functions associated with classes A and B, respectively; and h_A and h_B are the a priori probabilities of classes A and B, respectively. A decision boundary can be drawn where

$$f_A(X) = K f_B(X) \tag{2.5}$$

where

$$K = \frac{h_B l_B}{h_A l_A} \tag{2.6}$$

A Bayes decision rule developed in this way will asymptotically approach the Bayes optimal classifier with increasing sample size (Specht (1990)).

The model consists of four layers: the input layer, the pattern layer, the summation layer, and the output layer. The model used in this work uses a summation of spherical gaussian

basis functions centered at each training data point to approximate the underlying probability density function (PDF). The standard deviation σ (what Specht calls the smoothing parameter) of the gaussians and the cost/prior parameter K must be optimized in order to train the PNN; making this an eager learning model. The following equation is the PDF estimator kernel used in this work.

$$f_p(X) = \frac{1}{(2\pi)^{\frac{v}{2}}\,\sigma^v}\frac{1}{Ni_p}\sum_{i}^{Ni_p} exp\left[-\frac{(X - X_{pi})^t\,(X - X_{pi})}{2\sigma^2}\right] \tag{2.7}$$

In the above equation 2.7, f is the estimate of the underlying density function for the pth class; v is the number of variables; σ is the smoothing parameter; N is the number of training data points in the pth class; p is the class index; X is the test data point; and X_{pi} is the training data point. Equation 2.7 is the theoretical equation for the PNN. In practice, not all terms must be present for the performance to be optimal. Removing some terms will lighten the computational load. In the final analysis, all that is needed is a comparison among the probability density functions for each class. Like terms can be canceled out such as $1/\,(2\pi)^{\frac{v}{2}}\,\sigma^v$. Equation 2.8 shows the equation used in this work to apply the PNN.

$$f_p(X) = \frac{1}{Ni_p}\sum_{i}^{Ni_p} exp\left[-\frac{(X - X_{pi})^t\,(X - X_{pi})}{2\sigma^2}\right] \tag{2.8}$$

The structure of the PNN is given in figure 2.11. It shows a single test data point, two classes and eight data points in the training set. Each data point in the test set must be mapped to each data point in the training set using equation 2.8, summed according to the classes in the training set, and finally output an estimation of the probability that the test point is in each class. The class with highest probability is chosen as the estimated class of the test data point.

An advantage of the PNN is it can be trained very quickly. Only two parameters must be optimized for performance, after that, all the training data points must be kept on hand. The training process involves applying the PDF estimator kernel equation above for each class; making this a lazy learning model as the training is done just before classification. Highly multivariate data is easily handled. It is robust against bad or noisy data so long as the data

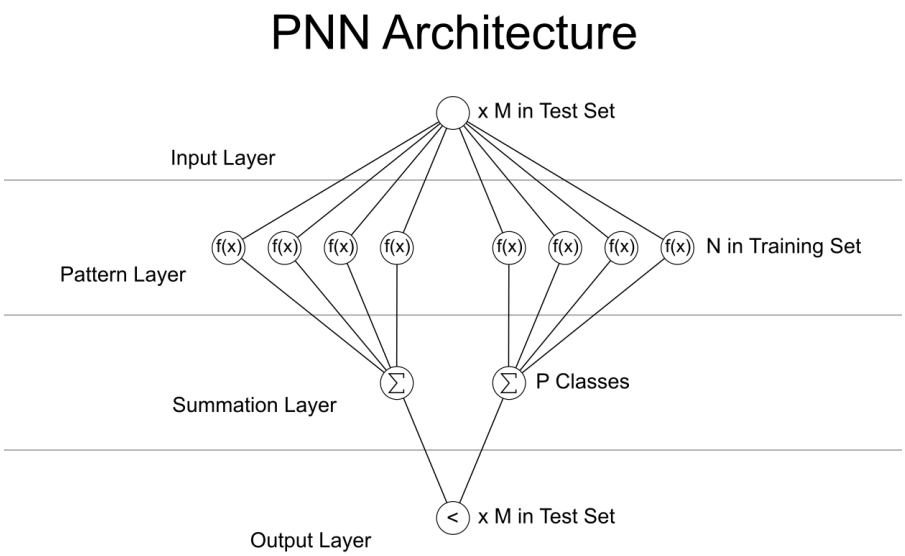
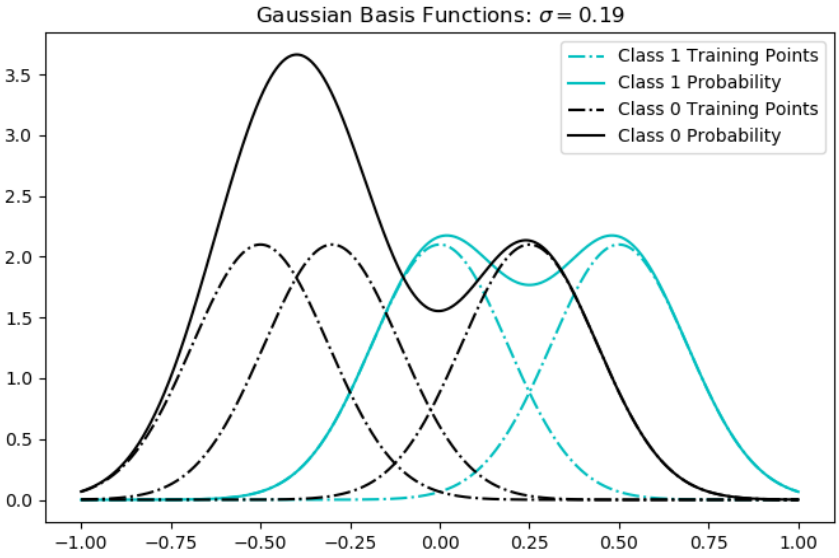


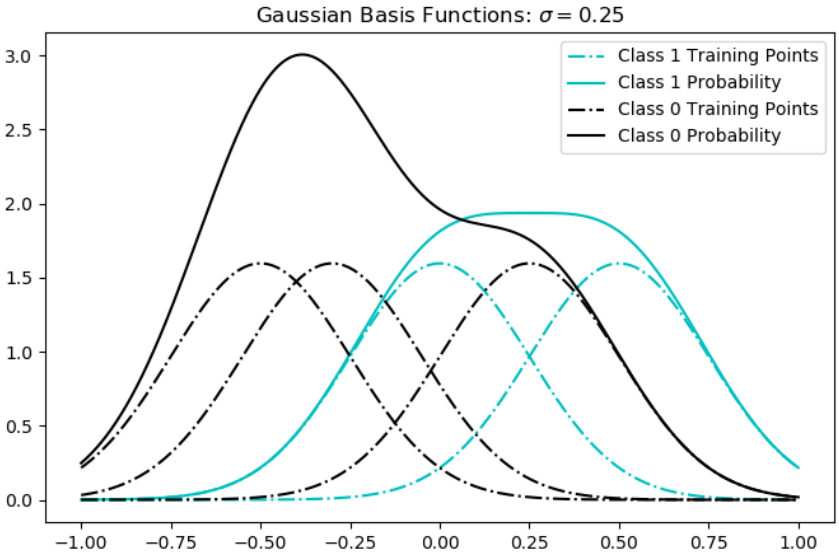
Figure 2.11: PNN model flow chart.

set is large. New data can be added quickly with or without retraining for a new σ value (Zaknich (2003)).

Choosing the correct σ value is critical to the performance of the PNN. Extremely low σ approaches a nearest neighbor rule, where the class is chosen based on the training point closest to the test point. Extremely high σ approaches a matched filter rule. Optimal σ has been proven to make the PNN approach a Bayes optimal classifier (Specht (1990)). Figure 2.12 shows an example of how σ affects the model for a data set of one dimension. The data set has three data points of class zero and two data points of class one. The result of the PNN is the max probability of a data point being part of a given class. If σ is 0.19, as in figure 2.12a , an unknown data point at $x = 0.25$ may be assigned class zero, but if σ is 0.25, as in figure 2.12b, it may be assigned class one. In this way, choice of σ can have a large impact on the behavior of the model.



(a) Training example with gaussian base standard deviation set to 0.19.



(b) Training example with gaussian base standard deviation set to 0.25.

Figure 2.12: Training example showing how the smoothing parameter σ affects the PNN.

2.7 Cross-Validation

Cross-validation is an important step in any machine learning application. A PNN requires cross validation to optimize parameters. Validation was performed in two ways in this work, leave-one-out cross-validation (LOO), and K-folds cross-validation.

LOO involves looping across all the data, setting aside one data point at a time, training on all the others, then testing with the one left out. Figure 2.13 illustrates the process. Each row represents one iteration of leave-one-out cross-validation. For each iteration, one data point is set aside and the rest are used for training. Spectra from the same kernel can't be used in both the training and testing sets while maintaining the validity of the model. For this reason, spectra from an entire seed were set aside at one time, thus speeding up the validation process. LOO is the simplest form of cross-validation. It gives the most optimistic estimate of the true performance of the model. LOO was used to optimize the σ and cost/prior weights of the model.

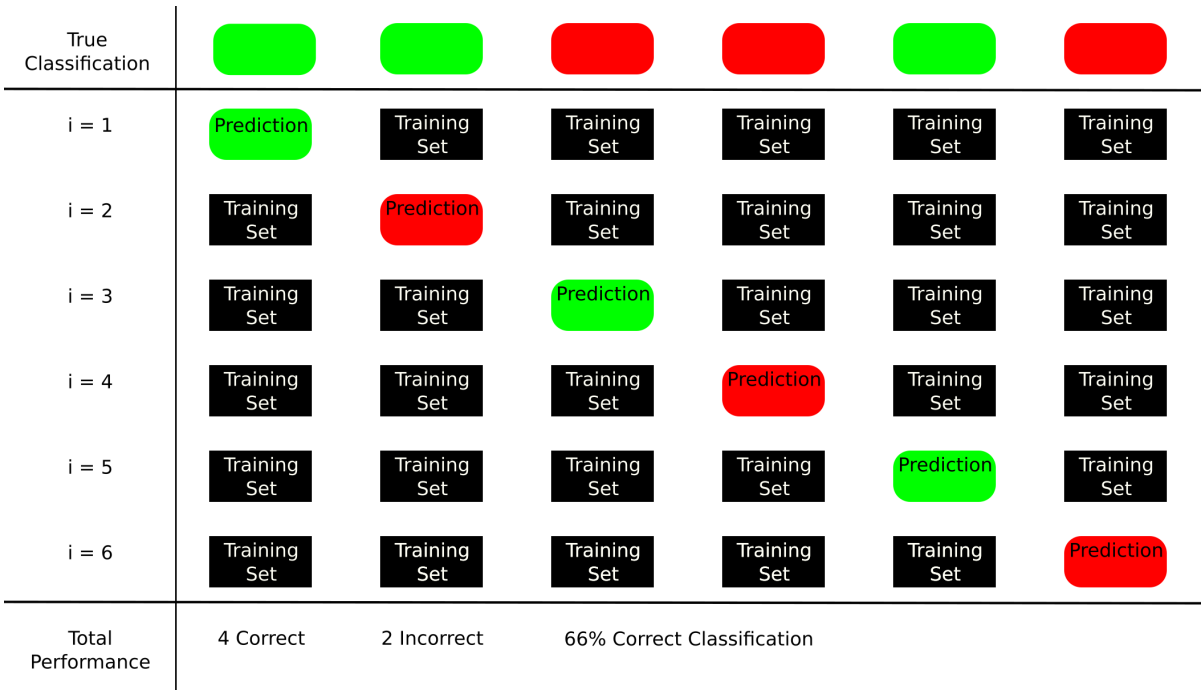


Figure 2.13: Example of leave-one-out cross-validation.

K-folds is a more trustworthy measure of real life performance. It involves separating the data into K groups, setting a group to be the test group and training on the rest, then rotating roles across the K groups. Figure 2.14 illustrates this process. Each row represents an iteration of K-folds cross-validation with three folds. For each iteration $\frac{1}{K}$ of the data set is set aside at random to be the test set, while the rest make up the training set. This must be done many times since the groups must be chosen randomly. In this work K-folds cross-validation was used as a confirmation tool to ensure the LOO cross-validation worked properly.

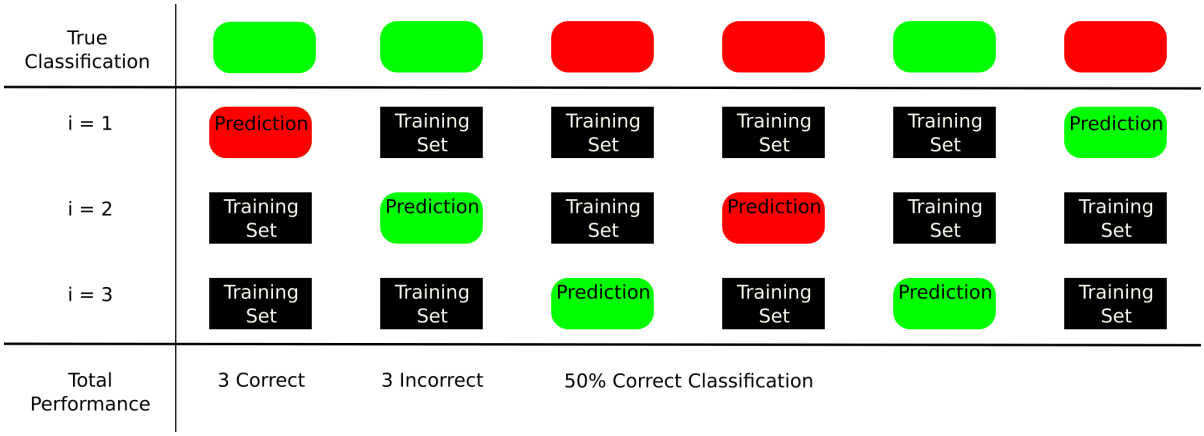


Figure 2.14: Example of K-Folds cross-validation, using three folds.

2.8 Training Subsampling

The last step in preprocessing the data set was choosing the right data to be in the training set. Model performance with the full set of data as the training set was not acceptable; as can be seen in table 3.2. Bolat and Yildirim (2003) used a random selection method for choosing the training set. Up to this point every spectrum in the data set has been used in training the model except those in the testing set. The method presented below will be a third group that will be used in testing like as before, but will never be used to train the model. The process is random in nature but is driven by the performance of the model. If there is a part of the data set that is negatively affecting the performance by polluting the training stage, this process

will remove it. In addition to the benefit in accuracy, this method as implemented reduces the training set by half, thus decreasing the time required to perform the training.

This subsampling methodology is an iterative approach using randomization to select the training set. The process has three steps:

1. Randomly assign each spectra a state that is either active or inactive. Spectra in the active state will participate in the training layer of the model, inactive spectra will not.

2. Choose N spectra at random and switch their states. If they were active previously, switch them to inactive, and vice versa.

3. Evaluate the performance of the new model using LOO cross-validation. If the performance improved, retain the N changes, if not, revert them and continue from step two.

For this work, N was set equal to 100. A high N value will result in more drastic changes in the model behavior, while a low N will not change the model significantly. Choice of N must be aligned with the difficulty of improving the model. If spectrum correct is the metric to gage improvement of the model, a low N can be used, and improvement can be gained slowly but incrementally. If kernel percent correct is used, a large N should be chosen because the model will require drastic changes in order to improve the kernel correct percentage. Both spectrum percent correct and total kernel percent correct were used as performance metrics in evaluating improvement in the model.

There is another more physical way of reducing the training set, and that is by using the topology of the corn kernel. The data set has time of flight information for each pixel, this is enough to tell how tall the corn kernel is at every location. spectra that are taken from locations where there is a high slope of kernel surface may be assumed to be sub-optimal, and thrown out. The remaining spectra in such a system would be from flat surfaces, leading to specular reflection. In the future this technique can be explored further. In this work the corn kernels had too much variance in height, and there was not enough scan resolution to make such a technique feasible.

2.9 PNN Software Structure

The PNN used in this work was programmed in Python. The training and testing data are passed to a PNN implementation function that tests any number of σ values. There are additional cross-validation functions that wrap the PNN implementation to do either LOO of K-folds cross-validation by passing training and testing data to the PNN function. Another function optimizes the K cost/prior values for each σ after the PNN has completed. Finally a results function processed the output of the PNN and evaluates its performance. The numerical library Numpy was used to accelerate the calculation through vectorization (Oliphant (2006)). Figure 2.15 shows a block diagram of how the code works.

2.10 PNN Validation and Behavior

A two dimensional data set was used to validate the code and explore the behavior of the PNN in a number of scenarios. The data source will be a 2D binary image of Yin and Yang. The image is square with 256 pixels on each side. The binary nature implies that there will be two classes in the data set, either zero or one. The data that will be used in the PNN is the X and Y position data. In figures 2.16 to 2.18 below, white area corresponds to class 0 and black area corresponds to class 1, red stars correspond to training set points within class 0 and green dots correspond to training set points within class 1. The black and white background is generated by applying the model using each pixel in the image as test data points. Such a complete testing set is not realistic, but comparing the results to the raw image may serve to give a sense of how the model will respond to changes in the training set and constitutive parameters. Model response to the following three variables will be explored in the rest of this section: Number of points in the training set, changes in σ, and noise in the training set.

The performance of the model improves as the training set size increases. This is an intuitive result, and easy to imagine. Figure 2.16 shows the improvement of performance as number of points increases. The distribution of the training data is important as well. Where the data is dense the model will work well, but it will work poorly where there are no training points. For this reason, a PNN is not useful for extrapolation.

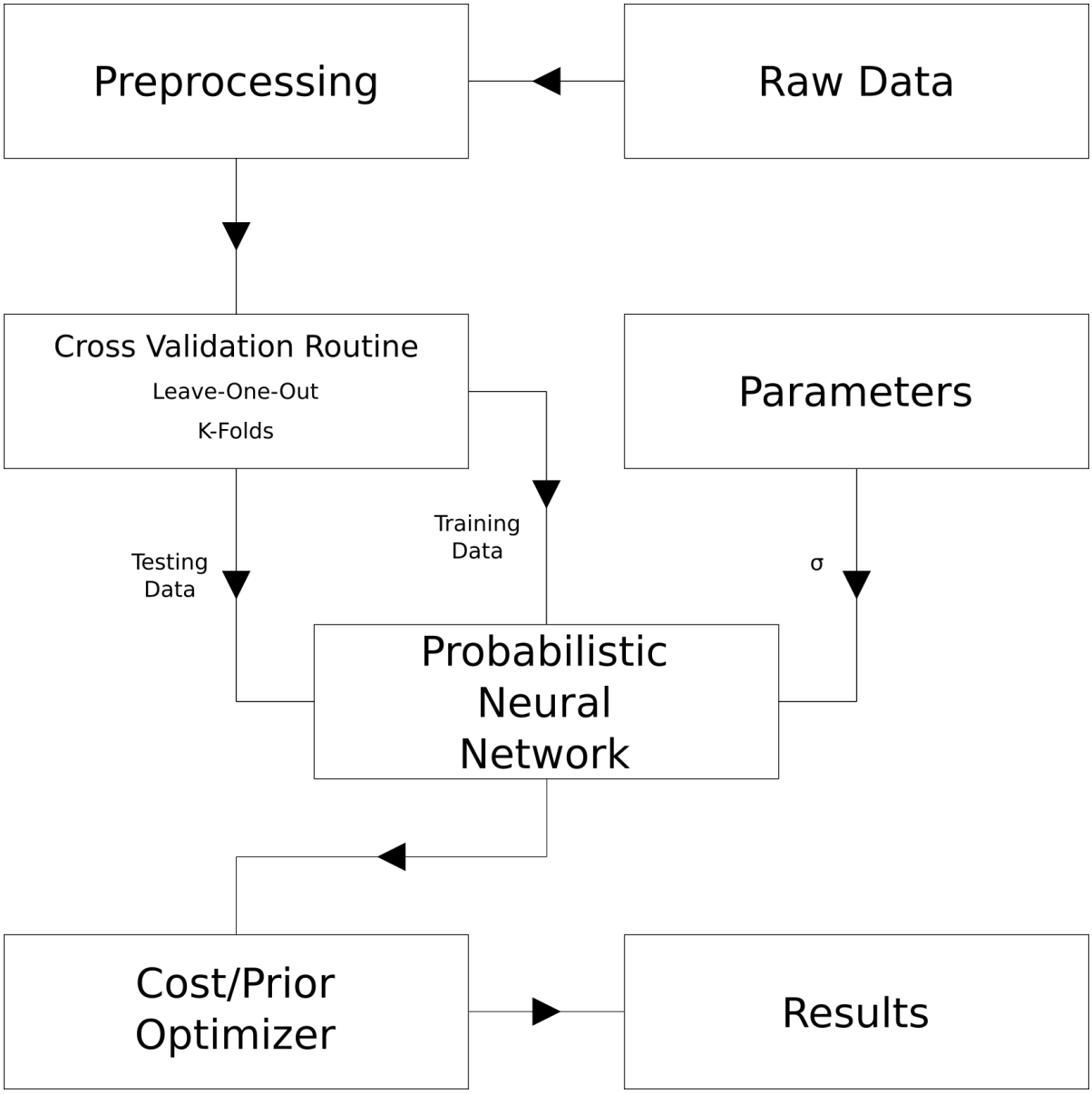


Figure 2.15: Program implementation diagram.

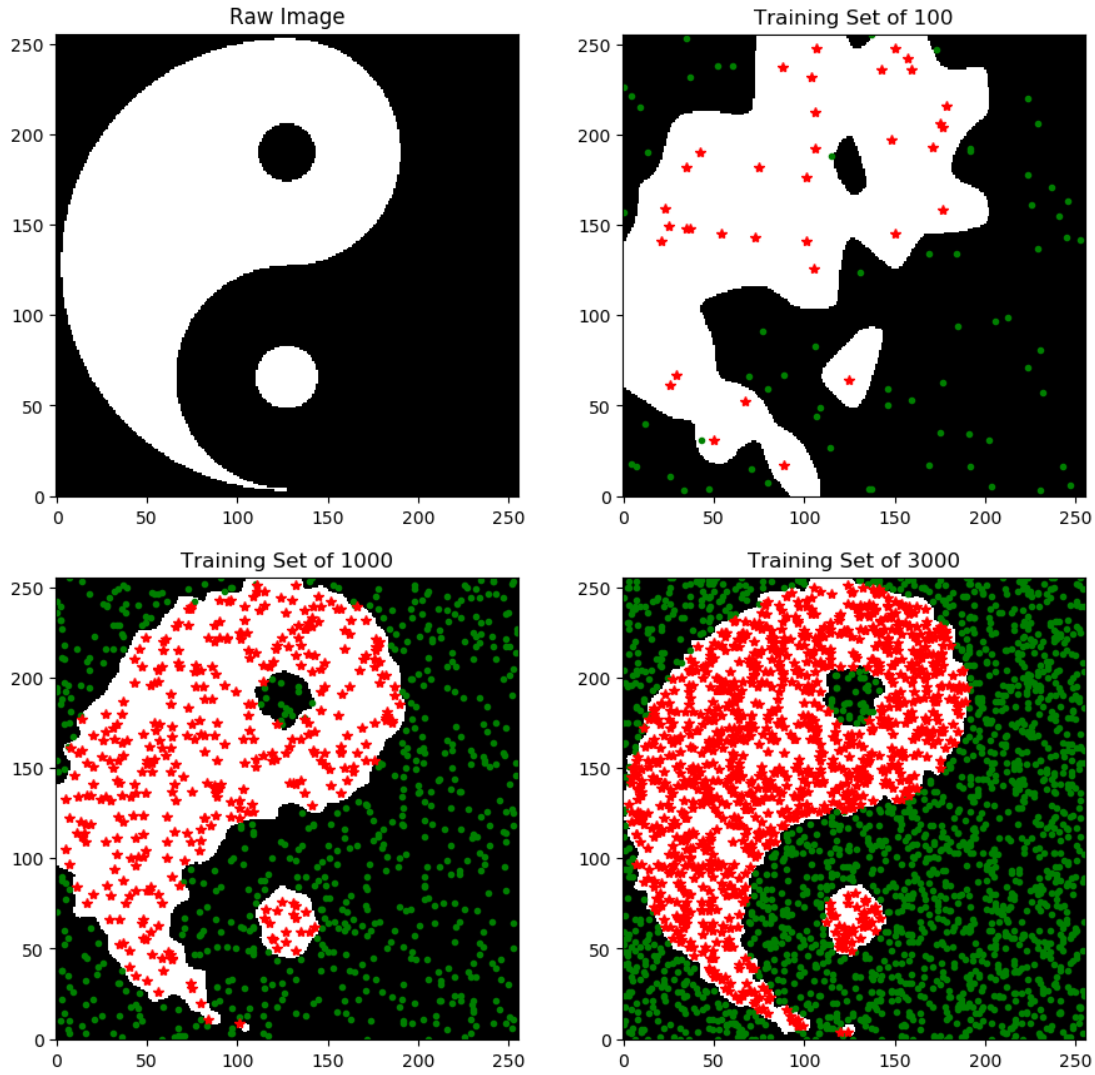


Figure 2.16: The performance of the model increases as the number of training points increases.

Figure 2.12 previously showed how σ can affect the model in a 1D test case. Figure 2.17 shows how σ affects the results on this 2D case. Extremely low σ approaches the nearest-neighbor rule. This behavior is observed.

Noise can pollute the training and lead to poor performance. Here, noise is added to the training set in the form of a percentage of points that are incorrectly classified. In the data shown, 10% of the training points are incorrect, adding noise to the training set. The PNN is robust against this noise, however, since a change in the σ can reduce the effects of noise. Figure 2.18 shows how the noise affects PNN performance, and how optimizing σ can reduce

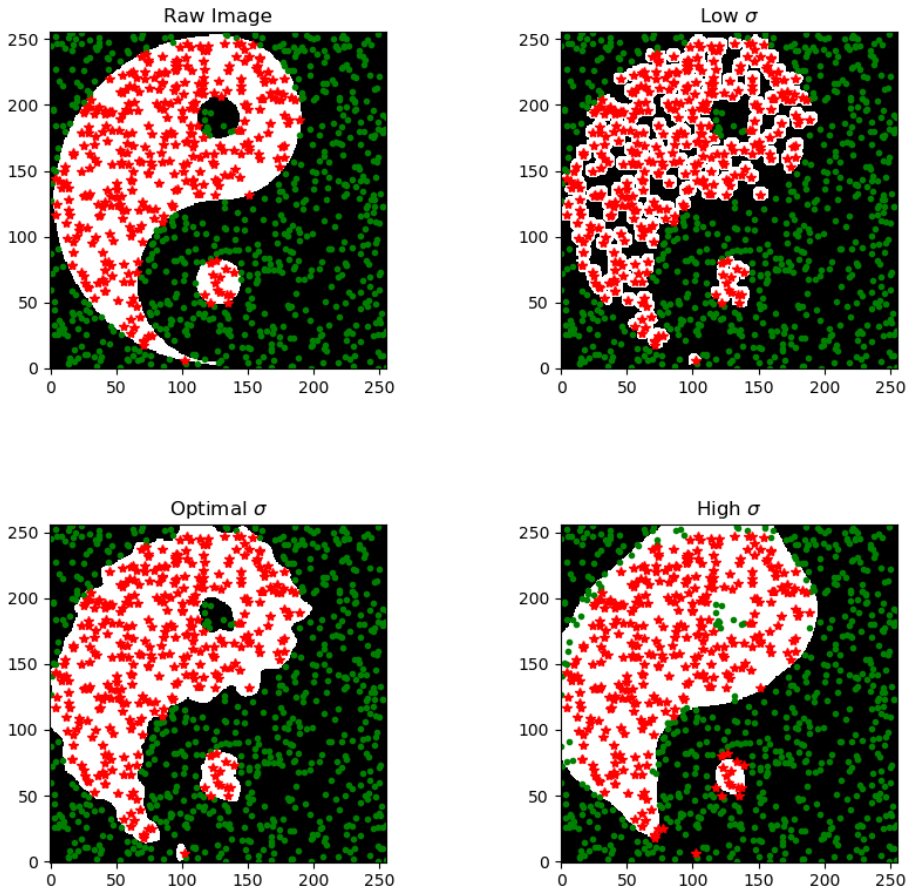


Figure 2.17: The value of σ has a large effect on PNN performance.

the effect. Subsampling the training set according to the technique described in section 2.8 is a good way to remove the noise in the training set.

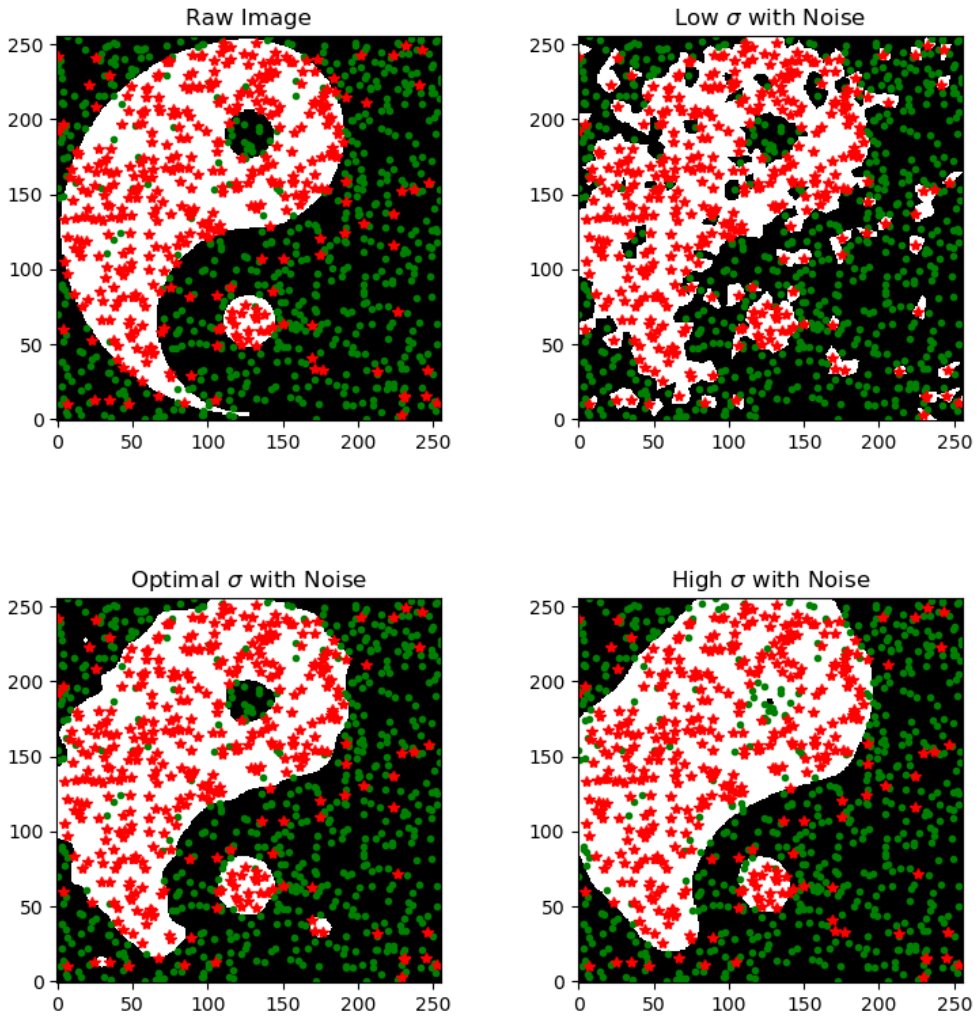


Figure 2.18: Noise in the training set has a small effect on the PNN performance.

CHAPTER 3. RESULTS

A preliminary scan of a flattened corn kernel was done to estimate properties such as the refractive index and the scan volume of a general kernel sample used in this study. Later, the PNN was applied to the data set on a per spectrum basis. Each spectra was identified according to the class that had the highest probability according to the PNN. Each seed was classified according to the majority class that its constituent spectra were identified. Two PNN models were built to differentiate haploid and diploid corn kernels using the data from the THz-TDS scan. One uses a σ smoothing value of 4.1238e-5 on spectra limited to between 0.0 and 1.0 THz, and the other uses a σ smoothing value of 1.3665e-4 on spectra limited to between 0.0 and 0.5 THz. Extensive cross-validation was performed on both models including LOO cross-validation to evaluate accuracy, and K-Folds cross-validation to evaluate robustness.

3.1 Preliminary Data Collection

Key properties of corn kernels with regard to this study include the refractive index and scan volume. A kernel was ground down to produce two flat, parallel sides with a thickness of approximately 1.55 mm (figure 3.2). This was accomplished using a Dremel tool and file. With this new geometry, a scan can pass through the kernel coherently and the bottom of the kernel can be imaged clearly.

An analysis of the time of flight between front surface and back surface echoes reveals the speed of light in the material ($\nu = 0.134\frac{\text{mm}}{\text{ps}}$), and thereby the refractive index of the material ($n = 2.238$), where c is the speed of light in a vacuum (equation 3.1). Figure 3.3 shows a representative waveform taken from the flattened kernel. The front surface is clear at about 85 ps, and the back surface lies at about 108 ps.

Figure 3.1: Corn kernel cross section.

The scan volume is here approximated by a cylinder of diameter equal to the spot size descending into the material. The length of the cylinder is determined by the attenuation of the kernel. At 1.55 mm the amplitude of the back surface is only 0.019, 16 times less than the front surface echo, and 100 times less than the reference waveform amplitude. In this scan, the scan volume is approximately 0.78 mm^3. At this depth, the signal is highly attenuated. Because this depth is much less than the height of the kernel, it can be assumed that most of the information gathered in this way will be contained in the front surface echo.

$$n = \frac{c}{\nu} \tag{3.1}$$

3.2 Frequency Band Selection

Using the full bandwidth provided by the waveform collected is unreasonable. Such a spectrum extends to as high as 12 THz. It is known that the photoconductive antennas used in these equipment have bandwidth up to 4.0 THz. A model was built to use just this part of the spectrum, but it failed to identify an effective smoothing factor. Similar models were tested all starting from 0.0 THz (dc) and extending to 3.5, 3.0, 2.5, 2.0, 1.5, 1.0, and 0.5 THz. Only the models built with 1.0 THz and 0.5 THz bandwidths showed promise, as can be seen in figures

Figure 3.2: Corn kernel ground down to create flat, parallel sides.

3.4 and 3.5, respectively, and also table 3.2. Figures 3.4 and 3.5 represent the performance of the PNN while optimizing the cost/prior weights for best spectra classification accuracy. There are choices of σ in each band that perform above 50% in each of the relevant metrics. These values of σ can be used to further train the model by other means, such as the subsampling technique.

This frequency range is low compared to other spectroscopic works in the THz range. Table 3.1 shows the frequency ranges of other spectroscopic research in the THz regime. Geometric variability of the corn kernel surface may be the cause of the poor performance of the higher frequency data. The high frequency waves are scattered off the surface and poor responses are collected. It is known that electromagnetic waves in this frequency range interact with collective vibrational and torsional modes in molecules. In the work by Sun et al. (2010), a tablet made of corn DNA was correctly classified using THz spectroscopy, this work may be sensing DNA as well. However, the specific modes leading to the classification are very hard to isolate due to the extreme variability of biological samples. It is believed that the choice and optimization of the frequency bands may relate to some physical properties of the corn kernels, but the underlying relationships are complicated and beyond the scope of this work.

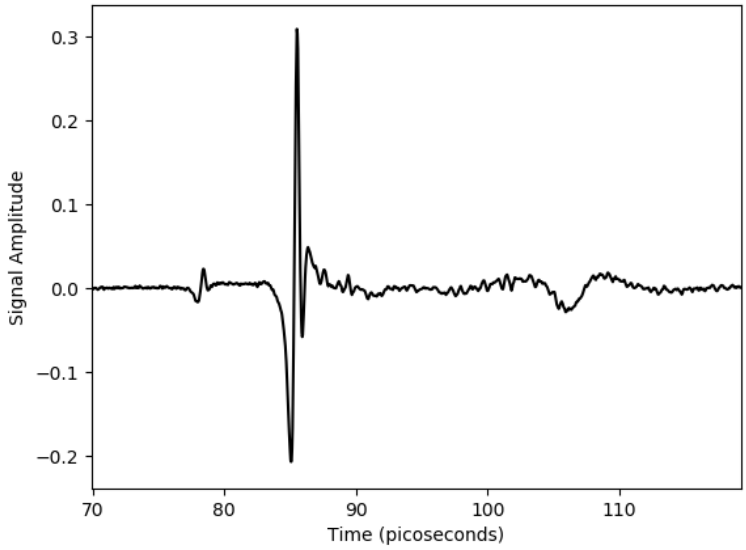


Figure 3.3: A-scan from the kernel after grinding.

Table 3.1: The frequency range of spectroscopic reasearch in the THz regime.

Source	**Frequency Band (THz)**	**Application**
Sun et al. (2010)	0.2-1.6	Classifying types of maize
Ashworth et al. (2009)	0.2-2.0	Classifying cancerous and healthy tissue
Yin et al. (2016)	1.5-3.5	Classifying edible oils
Lian et al. (2017)	0.2-1.6	Classification of transgenic maize

3.3 Training Set Optimization

Both models were optimized by reducing the training set size according to the method in section 2.8. This method was very successful using LOO cross-validation. Table 3.3 shows the results. As high as 31.9% improvement in kernel classification rate was observed. For the 0.5 THz bandwidth case the training set was reduced to 48.5% of its original size, figure 3.6 shows which pixels are used in the training stage.

This result was attained by running the optimizer for 12,656 iterations over about three days using a personal computer. Since it is a purely random phenomenon, there is no guarantee that a global optimum has been reached. The optimizer was stopped when it failed to improve the

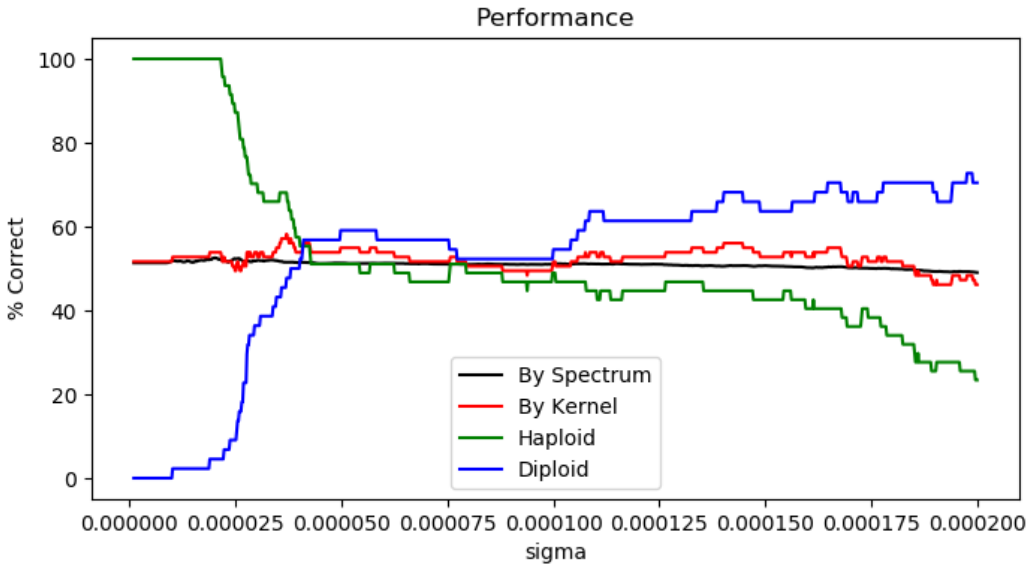


Figure 3.4: Leave-one-out cross-validation using the full training data set. The bandwidth here is between 0.0 and 1.0 Thz. Notice at $\sigma = 4.1238e-5$ where the haploid, diploid, and A-scan percent correct go above 50%.

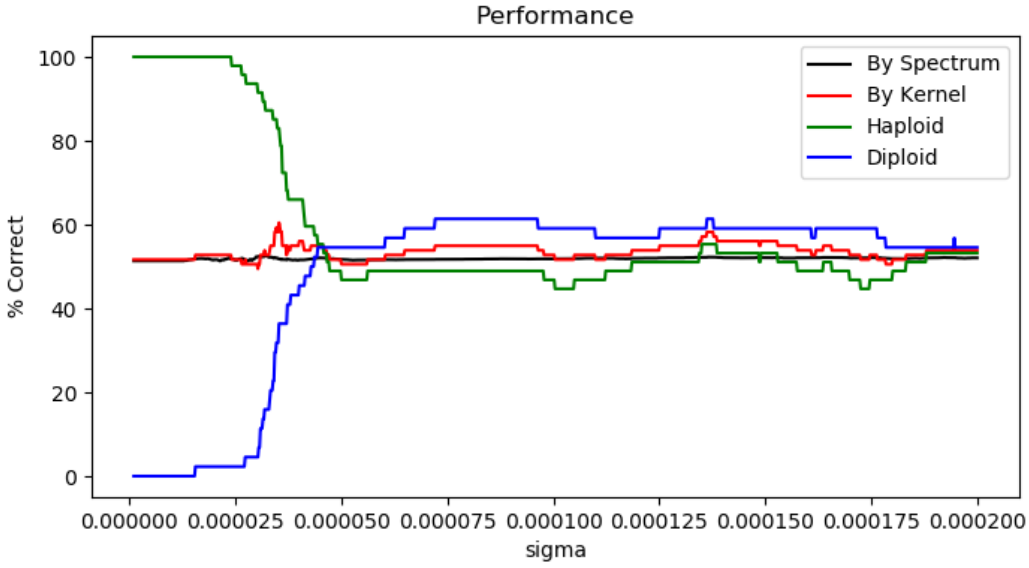


Figure 3.5: Leave-one-out cross-validation using the full training data set. The bandwidth here is between 0.0 and 0.5 Thz. Notice at $\sigma = 1.3665e-4$ where the haploid, diploid, and A-scan percent correct go above 50%.

Table 3.2: Leave-one-out cross-validation results before training set reduction, with 0.5 and 1.0 THz bandwidths.

Performance Marker	**0-0.5 THz**	**0-1.0 THz**
Smoothing Factor	1.3665e-4	4.1238e-5
Spectrum Accuracy	52.24 (%)	51.4 (%)
Haploid Accuracy	55.32 (%)	55.3 (%)
Diploid Accuracy	61.36 (%)	56.8 (%)
Kernel Accuracy	58.24 (%)	56 (%)

Table 3.3: Leave-one-out cross-validation results after reducing the training set, with 0.5 and 1.0 THz bandwidths.

Performance Marker	**0-0.5 THz**	**0-1.0 THz**
Smoothing Factor	1.3665e-4	4.1238e-5
Best Spectrum Accuracy	61.5 (%) (+9.26)	59.6 (%) (+8.2)
Best Haploid Accuracy	87.2 (%) (+31.88)	85.1 (%) (+29.8)
Best Diploid Accuracy	86.4 (%) (+25.04)	90.9 (%) (+34.1)
Best Kernel Accuracy	86.8 (%) (+28.56)	87.9 (%) (+31.9)

model after iterating for a long period of time, approximately 12 hours. A possible improvement to this optimizer would be to implement some evolutionary optimization methods, making use of a population of candidates and using crosses and random mutation to vary the training set.

3.4 Classification Robustness

K-folds cross-validation was used for both models. Good performance in K-folds cross-validation with few folds is a sign of robust models. Both models exhibit good performance in K-folds using 5 folds. Table 3.4 shows some statistics of the procedure using 5 folds, as well as figures 3.7 and 3.8.

K-folds cross-validation was performed with 7, 10, 13, 15, and 20 folds as well. Each increase in number of folds means smaller group sizes, better approximating the LOO cross validation. K-fold cross validation at 91 folds is equivalent to LOO in this data set. As expected, the performance increases steadily as the number of folds increases, simulating an increase in training data set size. Figure 3.9 shows how the performance increases as number of folds

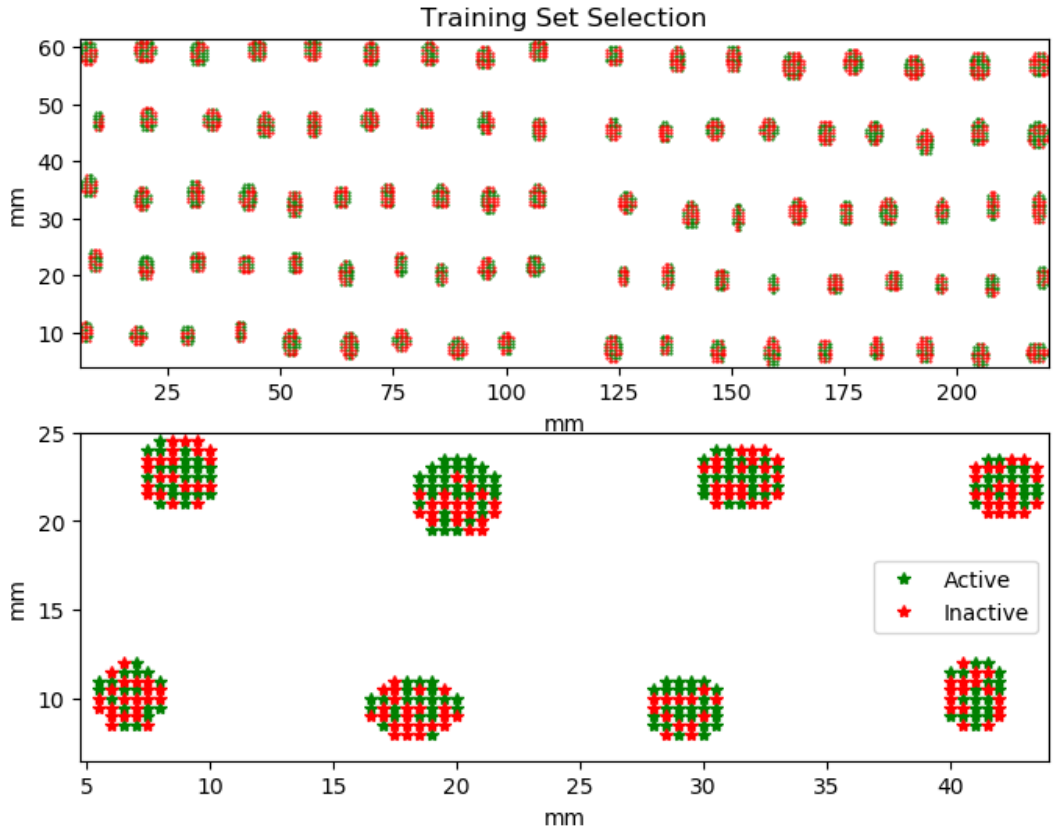


Figure 3.6: The training set used for 0.5 THz bandwidth case.

increases. Tables 3.5 through 3.9 show the statistics for these other K-folds iterations, and figures 3.10 through 3.19 show boxplots representing them.

It should be noted that the spectrum classification was relatively poor. Only 59% of the spectra were classified correctly in the 0.5 THz bandwidth case. Because each kernel is identified according to the class that more than 50% of its constituent spectra are identified, the relatively low classification accuracy of a single spectrum can be combined with other spectra on the same kernel. In this way the model can perform much better on each kernel as a whole because each kernel has many spectra. A higher spectrum performance would lead to more robust models and would require less data per kernel in the training set.

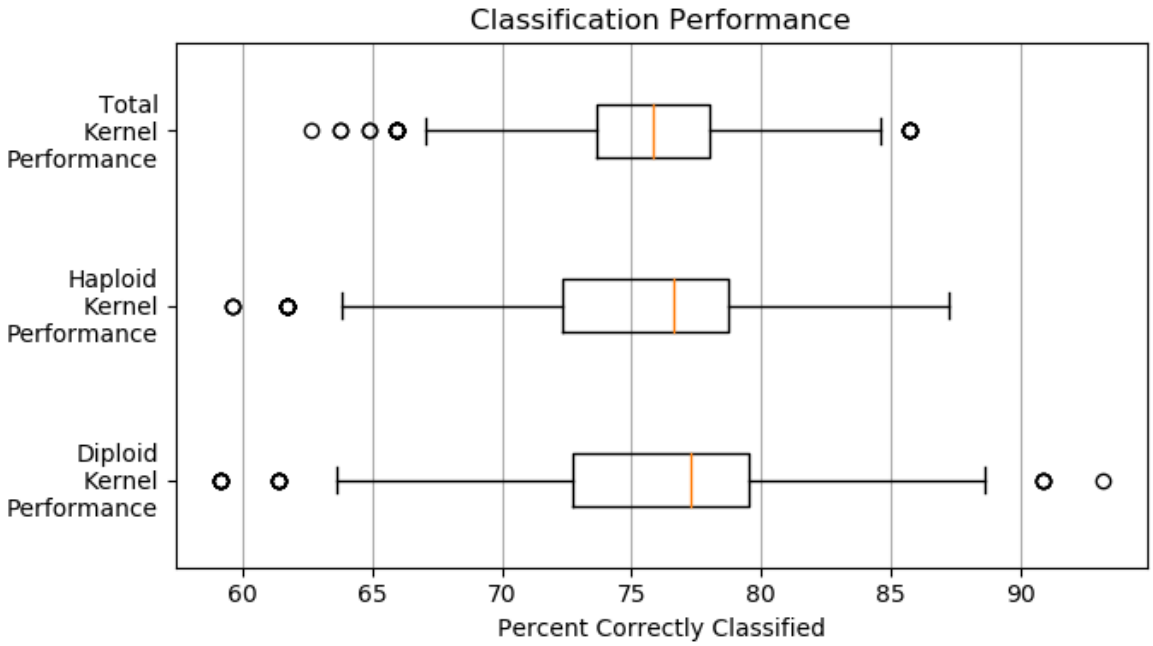


Figure 3.7: K-folds cross-validation results with 5 folds, using the 0.0-0.5 THz band.

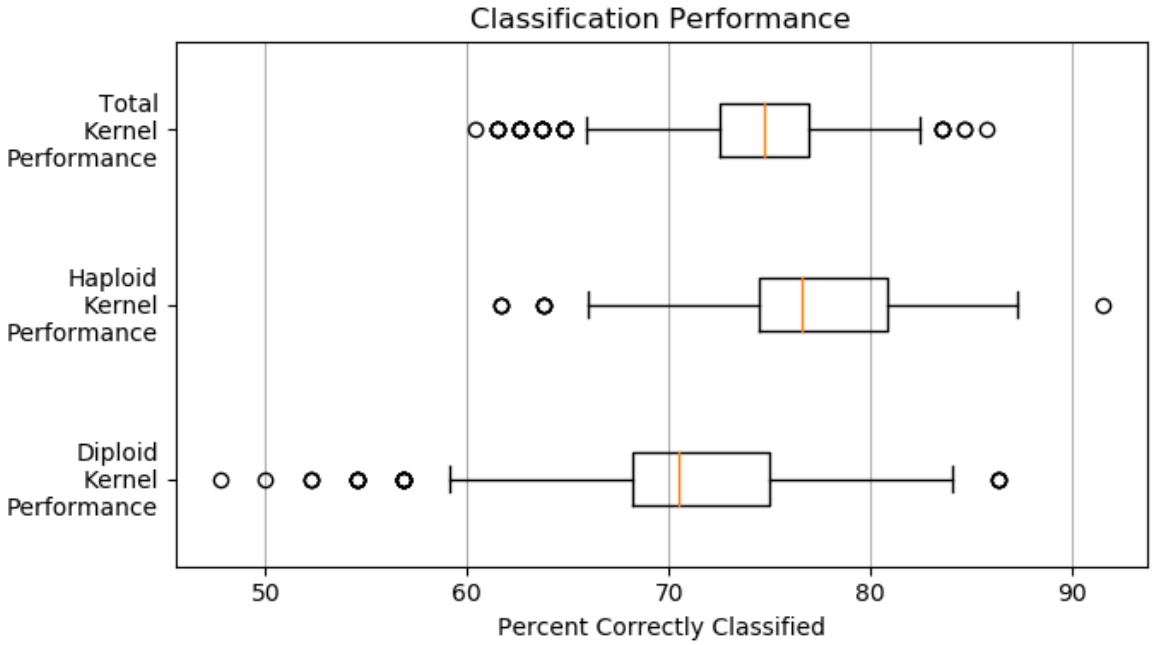


Figure 3.8: K-folds cross-validation results with 5 folds, using the 0.0-1.0 THz band.

Table 3.4: 5-Fold cross-validation results using the subsampled training data set with bandwidths of 0.5 and 1.0 THz.

5-Folds		**Spectrum (%)**	**Kernel (%)**	**Haploid (%)**	**Diploid (%)**
0.0-0.5 THz	**Median**	59.29	75.8	76.6	77.3
	S.D.	0.84	3.56	4.67	5.05
$N = 2448$	**Variance**	0.71	12.67	21.81	25.5
0.0-1.0 THz	**Median**	57.24	74.72	76.6	70.45
	S.D.	0.73	3.65	3.98	5.69
$N = 1632$	**Variance**	0.53	13.32	15.84	32.38

Table 3.5: 7-Fold cross-validation results using the subsampled training data set with bandwidths of 0.5 and 1.0 THz.

7-Folds		**Spectrum (%)**	**Kernel (%)**	**Haploid (%)**	**Diploid (%)**
0.0-0.5 THz	**Median**	60.0	79.1	78.7	79.55
	S.D.	0.72	3.03	4.0	4.23
$N = 2427$	**Variance**	0.52	9.21	16.0	17.93
0.0-1.0 THz	**Median**	57.87	76.92	78.82	75.0
	S.D.	0.58	2.99	3.30	4.84
$N = 1642$	**Variance**	0.34	8.97	10.89	23.4

3.5 Comparison with Prior Work

When comparing this method of kernel classification to others, care must be taken regarding the use-case. Some competing methods are less resource intensive in terms of time and money but are less accurate, such as the weight based method developed by Smelser et al. (2015). These methods can serve a role in enriching the number of haploid kernels in a batch, reducing the workload of a later technician to inspect visually. Methods that boast higher accuracy may negate the need for the visual inspection step, such as those developed by Jones et al. (2012), Boote et al. (2016), and Fuente et al. (2017). The methodology presented in this work may be more effective as an enrichment tool.

Other factors must be mentioned when discussing the efficacy of THz-TDS for corn discrimination. For example, it takes a very long time to scan each kernel. Given 30 minutes for each scan, four scans total, and 91 seeds, each kernel would take about 80 seconds to complete.

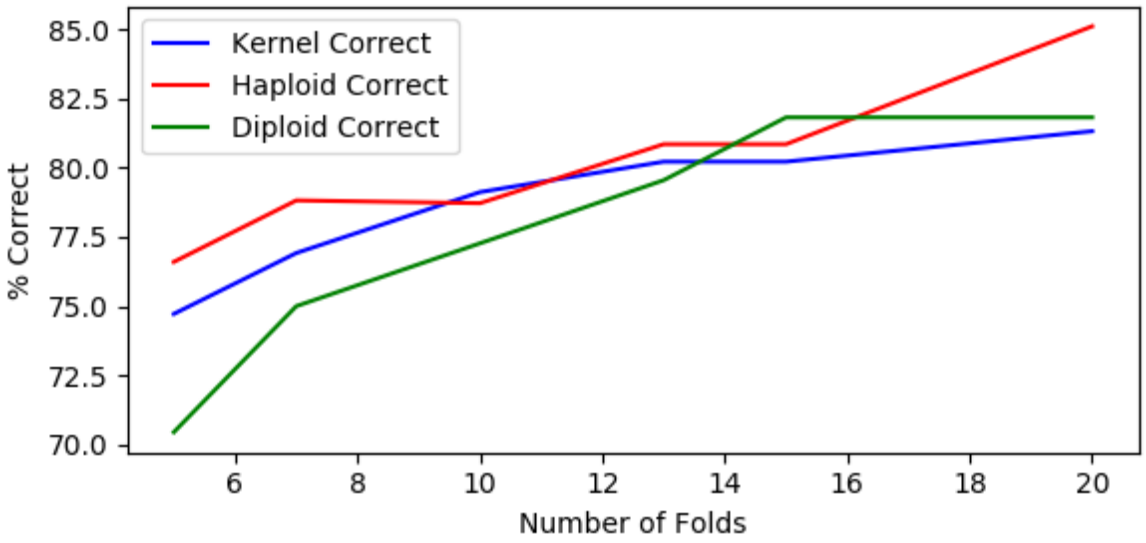


Figure 3.9: Performance progression with increasing number of folds in K-folds cross-validation.

A skilled technician can distinguish a haploid from a diploid corn seed in less than a second. Price should be taken into account as well. Today a THz-TDS system can cost hundreds of thousands of dollars. However, the technology is getting cheaper and faster as time passes. Future work could include exploring different ways of scanning data to improve throughput, such as including a conveyor belt. Perhaps a more precise location on the kernel can be found that shows variability that correlates with the haploid or diploid nature of the kernel. Scanning only this smaller area would increase efficiency and precision.

Table 3.6: 10-Fold cross-validation results using the subsampled training data set with bandwidths of 0.5 and 1.0 THz.

10-Folds		**Spectrum (%)**	**Kernel (%)**	**Haploid (%)**	**Diploid (%)**
0.0-0.5 THz	**Median**	60.47	80.22	80.85	79.55
	S.D.	0.56	2.59	3.48	3.67
$N = 2387$	**Variance**	0.32	6.70	12.14	12.14
0.0-1.0 THz	**Median**	59.30	79.12	78.72	77.27
	S.D.	0.49	2.55	2.89	4.10
$N = 1615$	**Variance**	0.24	6.49	8.34	16.81

Table 3.7: 13-Fold cross-validation results using the subsampled training data set with bandwidths of 0.5 and 1.0 THz.

13-Folds		**Spectrum (%)**	**Kernel (%)**	**Haploid (%)**	**Diploid (%)**
0.0-0.5 THz	**Median**	60.75	81.32	82.98	81.81
	S.D.	0.49	2.34	3.15	3.44
$N = 2417$	**Variance**	0.24	5.48	9.92	11.81
0.0-1.0 THz	**Median**	58.54	80.22	80.85	79.55
	S.D.	0.41	2.29	2.63	3.76
$N = 1568$	**Variance**	0.17	5.26	6.94	14.14

Table 3.8: 15-Fold cross-validation results using the subsampled training data set with bandwidths of 0.5 and 1.0 THz.

15-Folds		**Spectrum (%)**	**Kernel (%)**	**Haploid (%)**	**Diploid (%)**
0.0-0.5 THz	**Median**	60.88	82.42	82.98	81.82
	S.D.	0.45	2.13	2.85	3.14
$N = 2390$	**Variance**	0.21	4.56	8.14	9.87
0.0-1.0 THz	**Median**	58.66	80.22	80.85	81.82
	S.D.	0.37	2.16	2.44	3.52
$N = 1606$	**Variance**	0.13	4.68	5.97	12.36

3.6 Future Work and Improvements

There are many ways this work could be extended in the future. The experiment should be repeated with a much larger sample size. 91 seeds is a very small sample compared to the size of the corn breeding industry. This work could be extended by investigating the use of THz-TDS on high oil haploid inducers. The resultant kernels have more substantive chemical differences than those seen with the R1-nj marker, leading to possible markers in the THz frequency regime. Another important consideration is there may be other variables of interest such as damage, water content, or oil content that THz may be effective in sensing. More work needs to be done to explore these opportunities.

Improvements can be made to the machine learning and analysis processes as well. For example, kernel moisture is well correlated with FIR absorption. Including kernel moisture in the data set could improve the PNN performance. Improvements can be made in the

Table 3.9: 20-Fold cross-validation results using the subsampled training data set with bandwidths of 0.5 and 1.0 THz.

20-Folds		**Spectrum (%)**	**Kernel (%)**	**Haploid (%)**	**Diploid (%)**
0.0-0.5 THz	**Median**	61.07	83.52	85.11	84.1
	S.D.	0.37	1.92	2.57	2.72
$N = 2390$	**Variance**	0.14	3.67	6.60	7.40
0.0-1.0 THz	**Median**	58.84	81.32	80.85	81.82
	S.D.	0.33	1.85	2.15	3.02
$N = 1570$	**Variance**	0.11	3.43	4.63	9.15

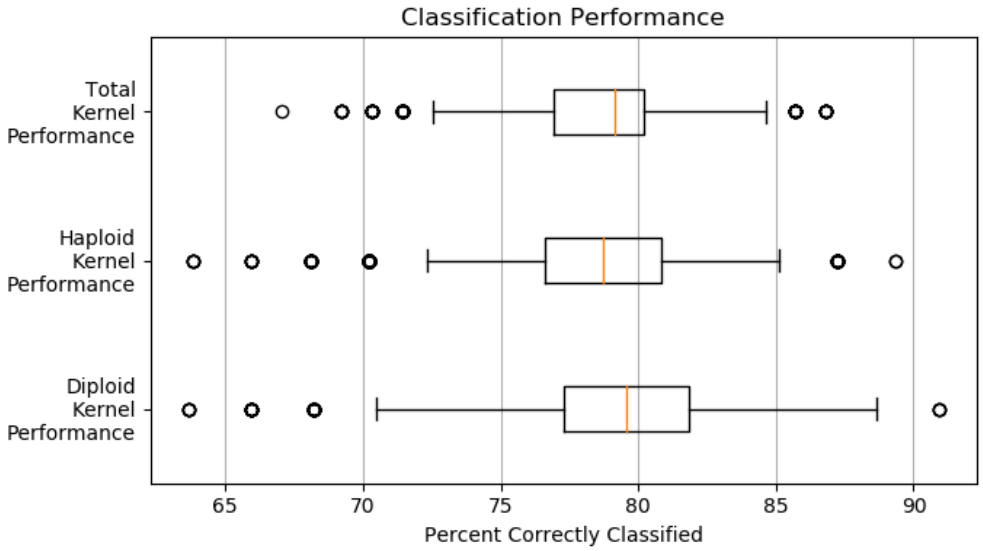


Figure 3.10: K-folds cross-validation results with 7 folds, using the 0-0.5 THz band.

optimization objective function as well. In this work the classification of both haploid and diploid kernels were optimized equally. A false negative (haploid declared diploid) is much worse than a false positive (diploid declared haploid) because of the relative scarcity of the haploids in the data set. It would be advantageous to reduce the false negative to zero at the expense of increasing the rate of false positives. The training subsampling methodology could be improved. It randomly assigns 50% of the training data to the active and inactive groups. Because the switching scheme is random, the active set doesn't stray far from 50% of total data. A more flexible subsampling technique should be implemented.

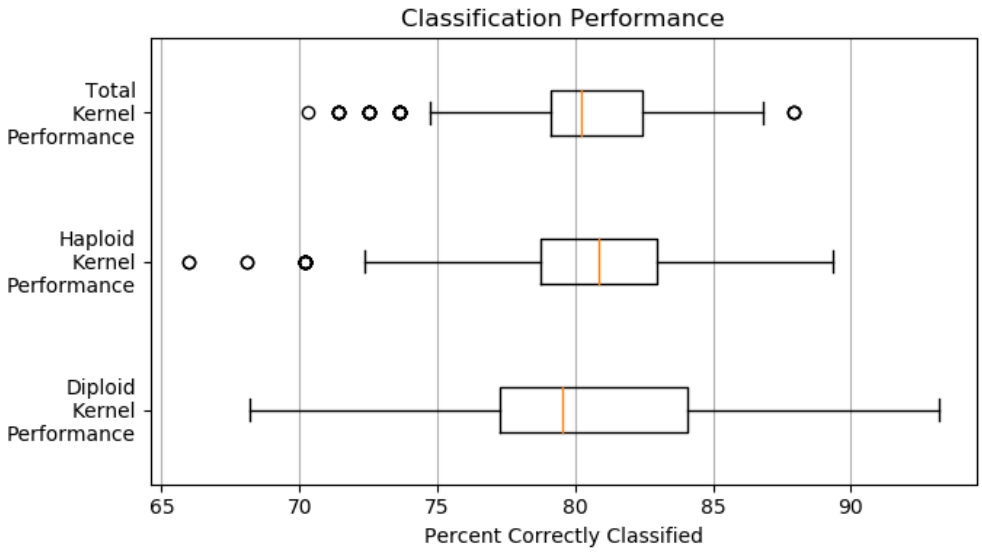


Figure 3.11: K-folds cross-validation results with 10 folds, using the 0-0.5 THz band.

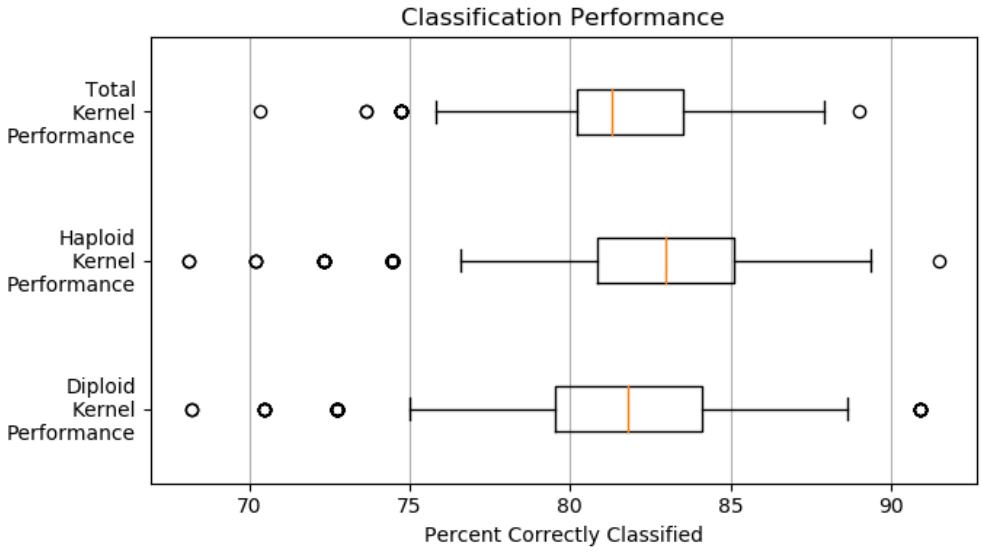


Figure 3.12: K-folds cross-validation results with 13 folds, using the 0-0.5 THz band.

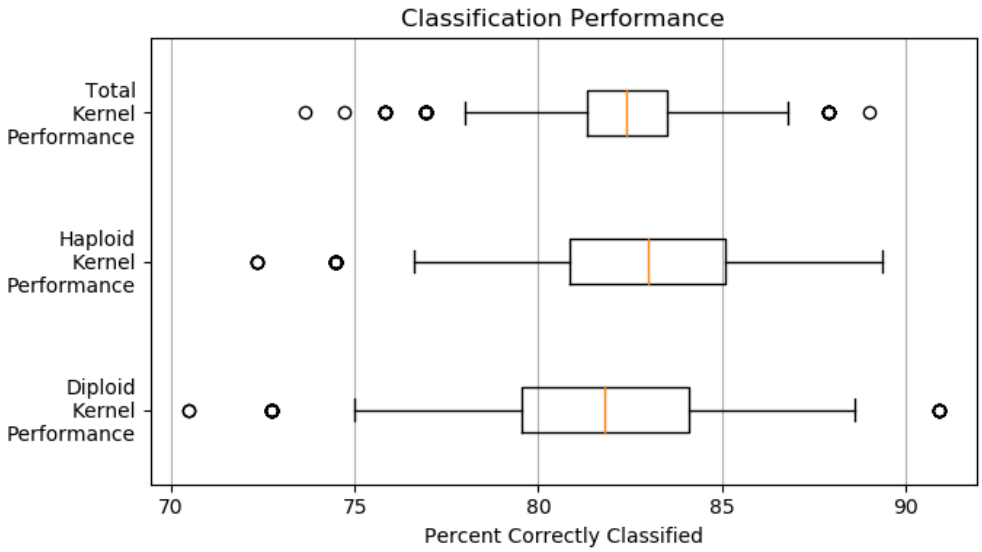


Figure 3.13: K-folds cross-validation results with 15 folds, using the 0-0.5 THz band.

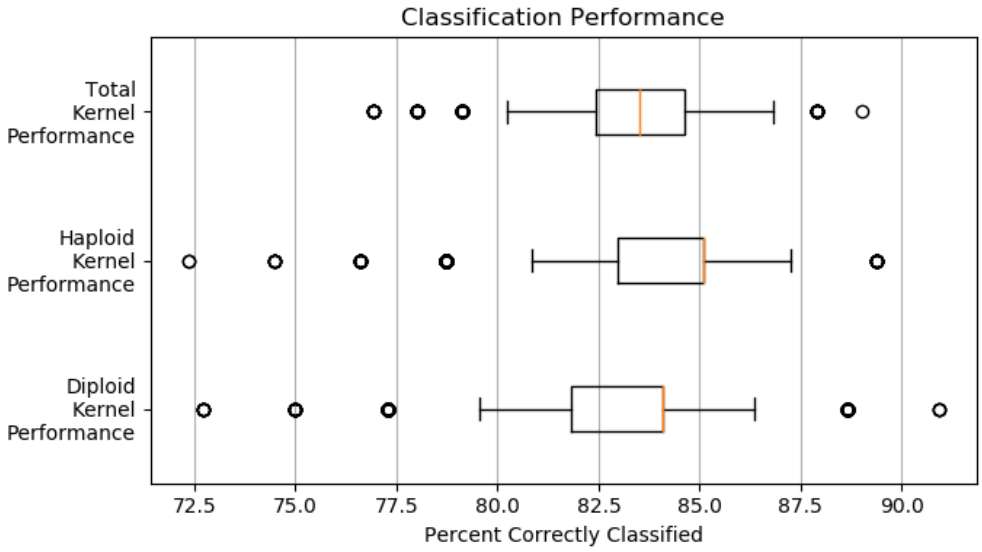


Figure 3.14: K-folds cross-validation results with 20 folds, using the 0-0.5 THz band.

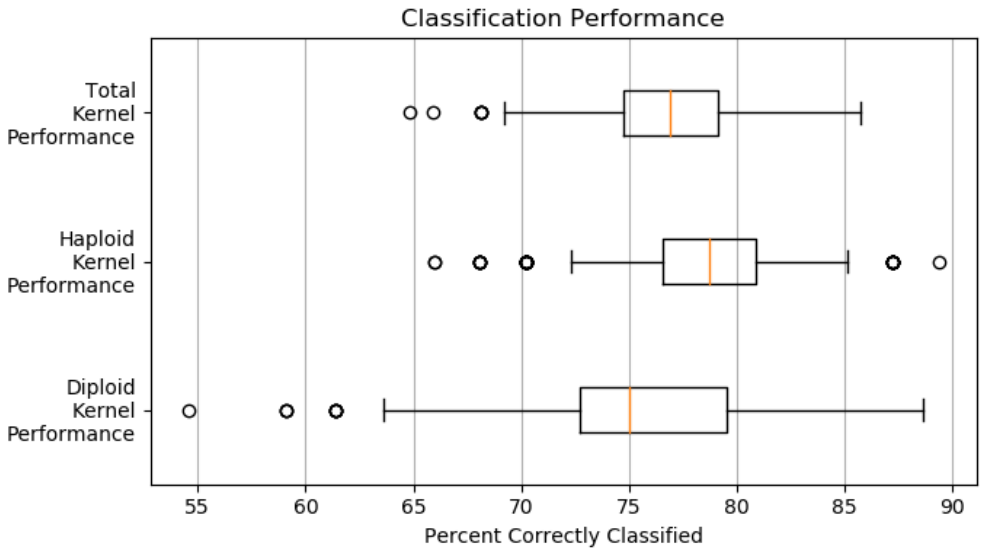


Figure 3.15: K-folds cross-validation results with 7 folds, using the 0-1.0 THz band.

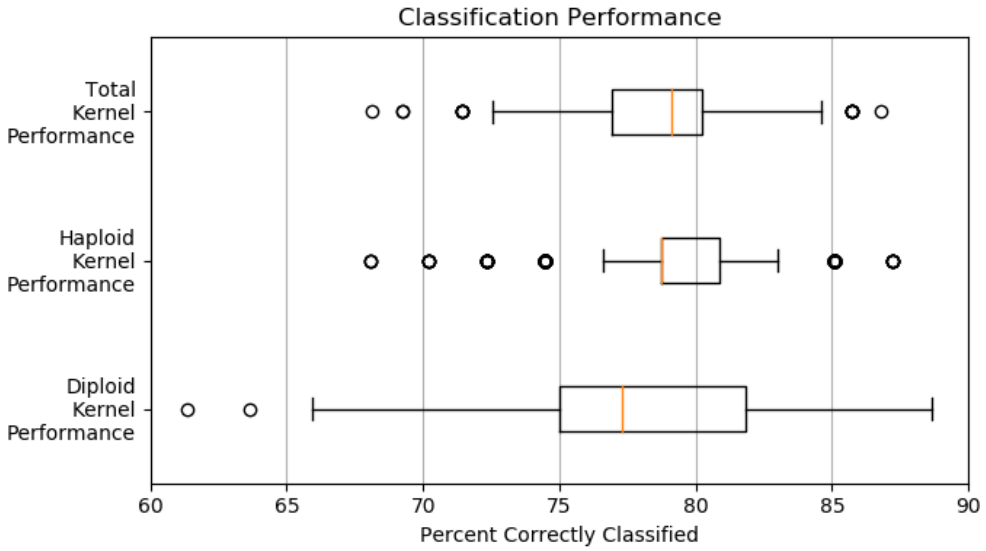


Figure 3.16: K-folds cross-validation results with 10 folds, using the 0-1.0 THz band.

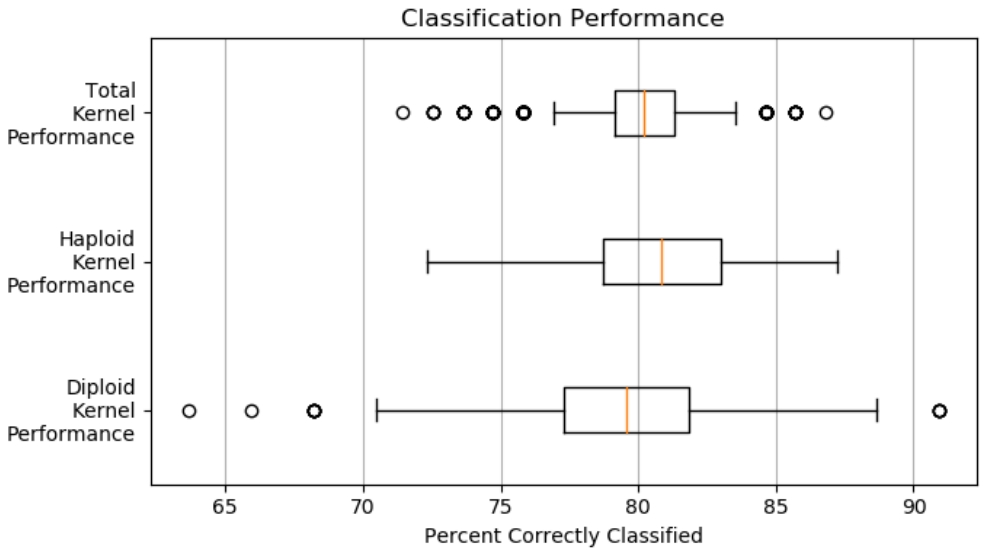


Figure 3.17: K-folds cross-validation results with 13 folds, using the 0-1.0 THz band.

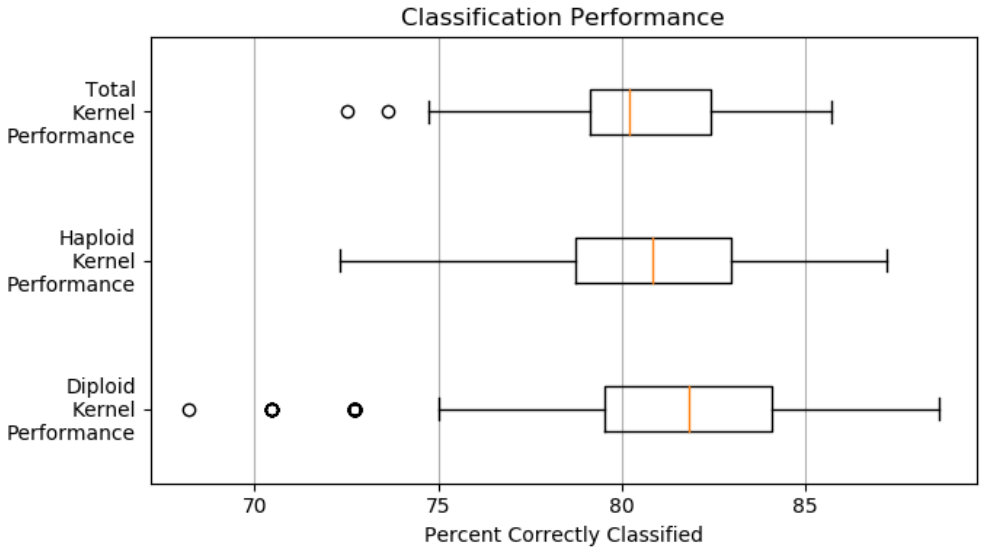


Figure 3.18: K-folds cross-validation results with 15 folds, using the 0-1.0 THz band.

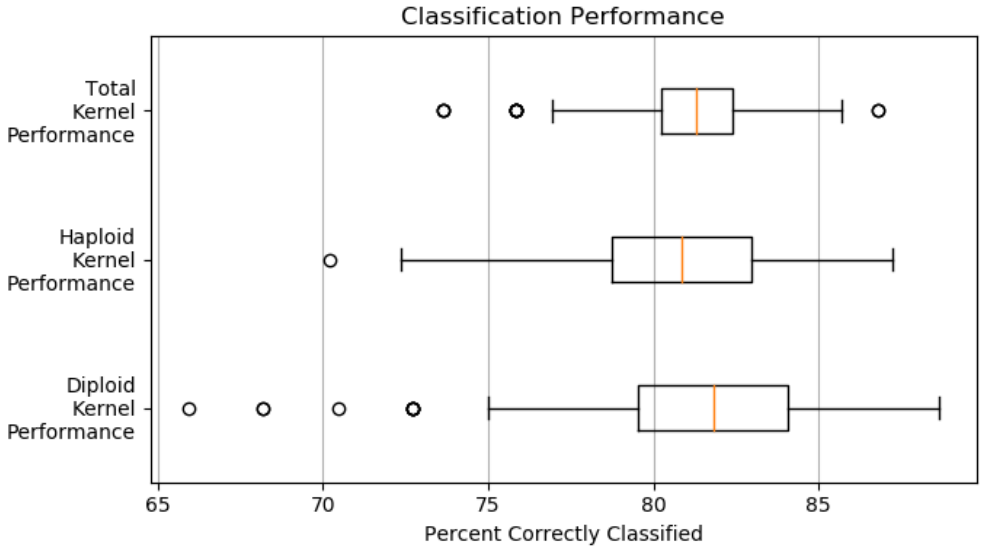


Figure 3.19: K-folds cross-validation results with 20 folds, using the 0-1.0 THz band.

CHAPTER 4. CONCLUSION

This work represents a novel approach to haploid and diploid classification. A Terahertz time-domain spectroscopy (THz-TDS) system was used to collect many waveforms per kernel. The data was transformed to frequency domain then analyzed using a probabilistic neural network (PNN). The PNN classified each spectrum and by proxy each kernel as either haploid or diploid. Leave-one-out cross-validation showed results as high as 87.9% for corn kernel classification. 5-fold cross-validation showed performance averaging around 75%, evidencing some stability in the model. These two results serve as a proof of principle that there may be an application for THz-TDS in classifying haploid and diploid corn kernels.

BIBLIOGRAPHY

Anastasi, R. and Madaras, E. (2006). Terahertz nde for under paint corrosion detection and evaluation. In Thompson, D. O. and Chimenti, D. E., editors, *Review of Progress in Quantitative Nondestructive Evaluation Volume 25*, volume 820, pages 515–522. AIP Conference Proceedings.

Ashworth, P. C., Pickwell-MacPherson, E., Provenzano, E., Pinder, S. E., Purushotham, A. D., Pepper, M., and Wallace, V. P. (2009). Terahertz pulsed spectroscopy of freshly excised human breast cancer. *Optics Express*, 17(15):12444–12454.

Banerjee, D., von Spiegel, W., Thomson, M. D., Schabel, S., and Roskos, H. G. (2008). Diagnosing water content in paper by terahertz radiation. *Optics Express*, 16(12):9060.

Bastke, S. (2009). Combining statistical network data, probabilistic neural networks and the computational power of GPUs for anomaly detection in computer networks. *Workshop Intelligent Security (SecArt 2009)*, (iii):1–6.

Boeing (2018). Advanced composite use.

Bolat, B. and Yildirim, T. (2003). Performance increasing methods for probabilistic neural networks. *Information Technology Journal*, 2(3):250–255.

Boote, B. W., Freppon, D. J., Fuente, G. N. D. L., Lübberstedt, T., Nikolau, B. J., and Smith, E. A. (2016). Haploid differentiation in maize kernels based on fluorescence imaging. *Plant Breeding*, 135(4):439–445.

Busch, S. F., Weidenbach, M., Fey, M., Schäfer, F., Probst, T., and Koch, M. (2014). Optical Properties of 3D Printable Plastics in the THz Regime and their Application for 3D Printed THz Optics. *Journal of Infrared, Millimeter, and Terahertz Waves*, 35(12):993–997.

Castro-Camus, E., Palomar, M., and Covarrubias, A. A. (2013). Leaf water dynamics of arabidopsis thaliana monitored in-vivo using terahertz time-domain spectroscopy. *Scientific Reports*, 3:1–5.

Catapano, I. and Soldovieri, F. (2017). A data processing chain for terahertz imaging and its use in artwork diagnostics. *Journal of Infrared, Millimeter, and Terahertz Waves*, 38(4):518–530.

Catapano, I., Soldovieri, F., Mazzola, L., and Toscano, C. (2017). Thz imaging as a method to detect defects of aeronautical coatings. *Journal of Infrared, Millimeter and Terahertz Waves*, 38(10):1264–1277.

Chan, W. L., Deibel, J., and Mittleman, D. M. (2007). Imaging with terahertz radiation. *Reports on Progress in Physics*, 70(8):1325–1379.

Chen, C.-C., Lee, D.-J., Pollock, T., and Whitaker, J. F. (2010). Pulsed-terahertz reflectometry for health monitoring of ceramic thermal barrier coatings. *Opt. Express*, 18(4):3477–3486.

Chiou, C. P., Margetan, F. J., Barnard, D. J., Hsu, D. K., Jensen, T., and Eisenmann, D. (2012). Nondestructive characterization of UHMWPE armor materials. In Thompson, D. O. and Chimenti, D. E., editors, *Review of Progress in Quantitative Nondestructive Evaluation*, volume 31, pages 1168–1175, Burlington, VT. American Institute of Physics (AIP), Conference Proceedings #1430.

Consolino, L., Bartalini, S., and De Natale, P. (2017). Terahertz frequency metrology for spectroscopic applications: A review. *Journal of Infrared, Millimeter, and Terahertz Waves*, 38(11):1289–1315.

De Lucia, F. (2003). Spectroscopy in the terahertz spectral region. In Mittleman, D., editor, *Sensing with Terahertz Radiation*, pages 39–115. Springer-Verlag Berlin.

Dorney, T. D., Baraniuk, R. G., and Mittleman, D. M. (2001). Material parameter estimation with terahertz time-domain spectroscopy. *J. Opt. Soc. Am. A*, 18(7):1562–1571.

Duvillaret, L., Garet, F., and Coutaz, J.-L. (1999). Highly precise determination of optical constants and sample thickness in terahertz time-domain spectroscopy. *Applied Optics*, 38(2):409.

Duvillaret, L., Garet, F., and Coutaz, J.-L. L. (1996). A reliable method for extraction of material parameters in terahertz time-domain spectroscopy. *IEEE Journal of Selected Topics in Quantum Electronics*, 2(3):739–746.

Fattinger, C. and Grischkowsky, D. (1989). Terahertz beams. *Applied Physics Letters*, 54(6):490–492.

Federici, J. F., Schulkin, B., Huang, F., Gary, D., Barat, R., Oliveira, F., and Zimdars, D. (2005). Thz imaging and sensing for security applications–explosives, weapons and drugs. *Semiconductor Science and Technology*, 20(7):S266.

Ferguson, B. and Zhang, X.-C. (2002). Materials for terahertz science and technology. *Nature Materials*, 1(1):26–33.

Food and Agricultural Organization (FAO) of the United Nations (2015). Food outlook: Biannual report of global food markets.

Fuente, G. N. D. L., Carstensen, J. M., Edberg, M. A., and Lübberstedt, T. (2017). Discrimination of haploid and diploid maize kernels via multispectral imaging. *Plant Breeding*, 136(1):50–60.

Fukuchi, T., Ozeki, T., Okada, M., and Fujii, T. (2016). Nondestructive inspection of thermal barrier coating of gas turbine high temperture components. *IEEJ Transactions on Electrical and Electronic Engineering*, 11(4):391–400.

Fukunaga, K. and Picollo, M. (2010). Terahertz spectroscopy applied to the analysis of artists' materials. *Applied Physics A: Materials Science and Processing*, 100(3):591–597.

Garoudja, E., Chouder, A., Kara, K., and Silvestre, S. (2017). An enhanced machine learning based approach for failures detection and diagnosis of pv systems. *Energy Conversion and Management*, 151:496–513.

Ge, H., Jiang, Y., Xu, Z., Lian, F., Zhang, Y., and Xia, S. (2014). Identification of wheat quality using thz spectrum. *Optics Express*, 22(10):12533.

Gente, R., Busch, S. F., Stubling, E. M., Schneider, L. M., Hirschmann, C. B., Balzer, J. C., and Koch, M. (2016). Quality control of sugar beet seeds with thz time-domain spectroscopy. *IEEE Transactions on Terahertz Science and Technology*, 6(5):754–756.

Hejase, J. A. (2012). *Terahertz time domain methods for material characterization of layered dielectric media*. Doctoral dissertation, Michigan State University.

Hilscher, E., Friedhoff, F., and Hirschmann, K. (2018). United States Patent Number 9,857,297 B2.

Hsu, D. K., Im, K. H., Chiou, C. P., and Barnard, D. J. (2011). An exploration of the utilities of terahertz waves for the NDE of composites. In Thompson, D. O. and Chimenti, D. E., editors, *Review of Progress in Quantitative Nondestructive Evaluation*, volume 30, pages 533–540, Burlington, VT. American Institute of Physics (AIP), Conference Proceedings #1335.

Hu, B. B. and Nuss, M. C. (1995). Imaging with terahertz waves. *Optics Letters*, 20(16):1716.

Jebarani, S. L. and Kamalaharidharini, T. (2017). Robust Face Recognition and Classification System Based on SIFT and DCP Techniques in Image Processing. In *International Conference on Electronics and Communication Systems*, volume 4, pages 43–48.

Jones, E., Oliphant, T., Peterson, P., et al. (2001–2017). SciPy: Open source scientific tools for python.

Jones, R. W., Reinot, T., Frei, U. K., Tseng, Y., Lübberstedt, T., and McClelland, J. F. (2012). Selection of haploid maize kernels from hybrid kernels for plant breeding using near-infrared spectroscopy and simca analysis. *Applied Spectroscopy*, 66(4):447–450.

Lee, Y.-S. (2009). *Principles of Terahertz Science and Technology*. Springer.

Lian, F., Xu, D., Fu, M., Ge, H., Jiang, Y., and Zhang, Y. (2017). Identification of Transgenic Ingredients in Maize Using Terahertz Spectra. *IEEE Transactions on Terahertz Science and Technology*, 7(4):378–384.

Liu, J. (2017). Terahertz spectroscopy and chemometrics classification of transgenic corn oil from corn edible oil. *Microwave and Optical Technology Letters*, 59(3):654–658.

Lopato, P. and Chady, T. (2013). Terahertz detection and identification of defects in layered polymer composites and composite coatings. *Nondestructive Testing and Evaluation*, 28(1):28–43.

Marrone, D. P., Blundell, R., Gibson, H., Paine, S., Papa, D. C., and Tong, E. (2004). Characterization and status of a terahertz telescope. In *15th International Symposium on Space Terahertz Technology*, number 2002, pages 426–432.

Mathanker, S. K., Weckler, P. R., and Wang, N. (2013). Terahertz (thz) applications in food and agriculture: a review. *Transactions of the ASABE*, 56(3):1213–1226.

Melchinger, A. E., Schipprack, W., Wuerschum, T., Chen, S., and Technow, F. (2013). Rapid and accurate identification of in vivo-induced haploid seeds based on oil content in maize. *Scientific Reports*, 3(2129).

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, New York.

Mittleman, D. (2003). Terahertz imaging. In Mittleman, D., editor, *Sensing with Terahertz Radiation*, pages 117–153. Springer-Verlag Berlin.

Mittleman, D. M. (2018). Twenty years of terahertz imaging [Invited]. *Optics Express*, 26(8):9417.

Neubert, P. and Protzel, P. (2014). Compact watershed and preemptive slic: On improving trade-offs of superpixel segmentation algorithms. In *2014 22nd International Conference on Pattern Recognition*, pages 996–1001.

Niijima, S., Shoyama, M., Murakami, K., and Kawase, K. (2018). Evaluation of the sintering properties of pottery bodies using terahertz time-domain spectroscopy. *Journal of Asian Ceramic Societies*, 6(1):37–42.

Oliphant, T. E. (2006). *A guide to numpy*. USA:Trelgol Publishing.

Palka, N., Panowicz, R., Ospald, F., and Beigang, R. (2015). 3D non-destructive imaging of punctures in polyethylene composite armor by tHz time domain spectroscopy. *Journal of Infrared, Millimeter, and Terahertz Waves*, 36(8):770–788.

Panwar, R. (2018). Performance and non-destructive evaluation methods of airborne radome and stealth structures. *Measurement Science and Technology*, 29(6):062001.

Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076.

Pickwell, E., Cole, B. E., Fitzgerald, A. J., Wallace, V. P., and Pepper, M. (2004). Simulation of terahertz pulse propagation in biological systems. *Applied Physics Letters*, 84(12):2190–2192.

Prasanna, B. M. (2012). Doubled haploid (dh) technology in maize breeding: An overview. In B. M. Prasanna, Vijay Chaikam, G. M., editor, *Doubled Haploid Technology in Maize Breeding: Theory and Practice*, chapter 1, pages 1–8. International Maize and Wheat Improvement Center.

Pupeza, I., Wilk, R., and Koch, M. (2007). Highly accurate optical material parameter determination with THz time-domain spectroscopy. *Optics Express*, 15(7):4335–4350.

Roeber, F. K., Gordillo, G. A., and Geiger, H. H. (2005). In vivo haploid induction in maize - performance of new inducers and significance of doubled haploid lines in hybrid breeding. *Maydica*, 50(3-4):275–283.

Rollwitz, W. L. (1989). Microwave inspection. In *Nondestructive Evaluation and Quality Control, ASM Handbook*. ASM International.

Shen, Y. C., Lo, T., Taday, P. F., Cole, B. E., Tribe, W. R., and Kemp, M. C. (2005). Detection and identification of explosives using terahertz pulsed spectroscopic imaging. *Applied Physics Letters*, 86(24):241116.

Siegel, P. (2002). Terahertz technology. *IEEE Transactions on Microwave Theory and Techniques*, 50(3):910–928.

Smelser, A., Blanco, M., Lübberstedt, T., Schechert, A., Vanous, A., and Gardner, C. (2015). Weighing in on a method to discriminate maize haploid from hybrid seed. *Plant Breeding*, 134(3):283–285.

Smith, P. R., Auston, D. H., and Nuss, M. C. (1988). Subpicosecond photoconducting dipole antennas. *IEEE Journal of Quantum Electronics*, 24(2):255–260.

Specht, D. F. (1990). Probabilistic neural networks. *Neural Networks*, 3(1):109–118.

Steenhoek, L. W. (1999). *A probabilistic neural network computer vision system for corn kernel damage evaluation*. PhD thesis.

Stoik, C., Bohn, M. J., and Blackshire, J. L. (2008). Nondestructive evaluation of aircraft composites using transmissive terahertz time domain spectroscopy. *Optics Express*, 16(21):17039,17051.

Suen, J. (2016). Terabit-per-second satellite links: A path toward ubiquitous terahertz communication. *Journal of Infrared, Millimeter, and Terahertz Waves*, 37(7):615–639.

Sun, J., Shen, J., Li, N., Lu, M., Jia, Y., Guo, J., and Zhang, J. (2010). Identifying type of maize with terahertz time-domain spectroscopy. In *International Conference on Biomedical Engineering and Informatics*, volume 3, pages 918–921.

Taylor, Z. D., Singh, R. S., Bennett, D. B., Tewari, P., Kealey, C. P., Bajwa, N., Culjat, M. O., Hubschman, J.-p., Brown, E. R., Grundfest, W. S., and Lee, H. (2011). THz medical imaging : in vivo hydration sensing. *IEEE Transactions on Terahertz Science and Technology*, 1(1):201–219.

van der Walt, S., Schnberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., and Yu, T. a. (2014). scikit-image: image processing in python. *PeerJ*, 2:e453.

Varmuza, K. and Filzmoser, P. (2009). *Introduction to multivariate statistical analysis in chemometrics*. CRC Press.

Verghese, S. (1997). Highly tunable fiber-coupled photomixers with coherent terahertz output power. *IEEE Transactions on Microwave Theory and Techniques*, 45(8 PART 2):1301–1309.

Vinodhini, G. and Chandrasekaran, R. M. (2016). A comparative performance evaluation of neural network based approach for sentiment classification of online reviews. *Journal of King Saud University - Computer and Information Sciences*, 28(1):2–12.

Walther, M., Fischer, B. M., Ortner, A., Bitzer, A., Thoman, A., and Helm, H. (2010). Chemical sensing and imaging with pulsed terahertz radiation. *Analytical and Bioanalytical Chemistry*, 397(3):1009–1017.

Wang, C., Qin, J. Y., Xu, W. D., Chen, M., Xie, L. J., and Ying, Y. B. (2018). Terahertz imaging applications in agriculture and food engineering: A review. *Transactions of the ASABE*, 61(2):411–424.

Wang, H., Liu, J., Xu, X., Huang, Q., Chen, S., Yang, P., Chen, S., and Song, Y. (2016). Fully-automated high-throughput nmr system for screening of haploid kernels of maize (corn) by measurement of oil content. *PLOS ONE*, 11(7):1–14.

Wang, K., Sun, D. W., and Pu, H. (2017). Emerging non-destructive terahertz spectroscopic imaging technique: Principle and applications in the agri-food industry. *Trends in Food Science and Technology*, 67:93–105.

Xu, W., Xie, L., Ye, Z., Gao, W., Yao, Y., Chen, M., Qin, J., and Ying, Y. (2015). Discrimination of transgenic rice containing the Cry1Ab protein using terahertz spectroscopy and chemometrics. *Scientific Reports*, 5(July):1–9.

Yakovlev, E. V., Zaytsev, K. I., Chernomyrdin, N. V., Gavdush, A. A., Zotov, A. K., Nikonovich, M. Y., and Yurchenko, S. O. (2016). Non-destructive testing of composite materials using terahertz time-domain spectroscopy. In *Proc. SPIE, Optical Sensing and Detection*, volume 9899.

Yin, M., Tang, S., and Tong, M. (2016). Identification of edible oils using terahertz spectroscopy combined with genetic algorithm and partial least squares discriminant analysis. *Analytical Methods*, 8(13):2794–2798.

Zaknich, A. (2003). *Neural Networks for Intelligent Signal Processing*, volume 4. World Scientific Publishing Co Pte Ltd, Singapore.

Zhang, X.-C. (2010). *Introduction to thz wave photonics*. Springer, New York.

Zhong, S. (2018). Progress in terahertz nondestructive testing: A review. *Frontiers of Mechanical Engineering*.

Zimdars, D., Valdmanis, J. A., White, J. S., Stuk, G., Williamson, S., Winfree, W. P., and Madaras, E. (2005). Technology and applications of terahertz imaging non-destructive examination: Inspection of space shuttle sprayed on foam insulation. In *AIP Conference Proceedings*, volume 760,570.